ADAPTIVE FEATURE FUSION FOR MULTI-CAMERA HUMAN TRACKING IN CROWDED ENVIRONMENTS

Mikhail Sorokin
m.sorokin@cvdtf.edu

Elizaveta Koroleva
e.koroleva@cvdtf.edu

Alexander Shuvalov
genshuvalov.alex@gmail.com

Artem Ilgamov
a.ilgamov@cvdtf.edu

ABSTRACT

Multi-camera human tracking in crowded environments remains a challenging problem due to occlusions, illumination changes, and appearance variations across different camera views. This paper presents a novel framework that leverages adaptive feature fusion and temporal consistency constraints to improve tracking performance in complex environments. Our approach combines appearance, motion, and spatial-temporal features through a dynamic weighting mechanism that adapts to the complexity of the scene. We introduce a Confidence-Aware Association (CAA) algorithm that explicitly models tracking uncertainty and uses it to guide the data association process. Extensive experiments on three public datasets demonstrate that our method achieves comparable or superior performance to state-of-the-art approaches, with notable improvements in crowded scenes with frequent occlusions. The proposed framework achieves a MOTA score of 76.8% on the WILDTRACK dataset and 68.3% on the CAMPUS dataset, representing a 2.6% and 3.1% improvement over baseline methods, respectively. Our ablation studies highlight the effectiveness of the adaptive feature fusion mechanism and temporal consistency constraints in improving tracking robustness and accuracy.

***K*eywords** Multi-camera tracking · pedestrian tracking · feature fusion · temporal consistency · multi-target tracking · computer vision

1 Introduction

Multi-camera human tracking is essential for various applications such as surveillance, sports analytics, crowd monitoring, and behavior analysis. By deploying multiple cameras with overlapping fields of view, tracking systems can overcome challenges like occlusions, illumination variations, and limited field of view that plague single-camera systems. However, effective integration of information from multiple cameras remains a significant challenge, particularly in crowded environments where occlusions and identity switches are common.

Traditional approaches to multi-camera tracking typically rely on camera calibration to establish geometric correspondence across views, followed by appearance matching to associate detections across cameras. Recent deep learning methods have shown promising results by learning discriminative appearance features directly from data. However, these approaches often struggle in crowded scenes where people with similar appearances are in close proximity, leading to identity switches and fragmented trajectories.

In this paper, we address these challenges by proposing OmniTrack, a framework that adaptively fuses multiple feature types based on scene conditions. Unlike existing methods that use fixed weighting schemes for feature fusion, our approach dynamically adjusts the importance of appearance, motion, and spatial-temporal features based on their reliability in the current context. This adaptive fusion strategy is particularly effective in challenging scenarios where certain features may be unreliable due to occlusions, lighting changes, or similar appearances.

The key contributions of our work are as follows:

1. We propose an adaptive feature fusion mechanism that dynamically adjusts feature weights based on their estimated reliability in the current scene context.

2. We introduce a Confidence-Aware Association (CAA) algorithm that explicitly models tracking uncertainty and uses it to guide the data association process.

3. We develop temporal consistency constraints that exploit the smooth dynamics of human motion to improve tracking continuity and reduce identity switches.

4. We conduct extensive experiments on three public datasets, demonstrating that our approach achieves competitive results compared to state-of-the-art methods, with particular improvements in crowded scenes.

The rest of the paper is organized as follows: Section 2 discusses related work in multi-camera tracking. Section 3 presents our proposed OmniTrack framework in detail. Section 4 describes our experimental setup and results. Section 5 provides ablation studies to analyze the contribution of each component. Finally, Section 7 concludes the paper and discusses future directions.

2 Related Work

2.1 Single-Camera Multi-Target Tracking

Single-camera multi-target tracking has seen significant advances with the emergence of tracking-by-detection paradigms [1, 2]. These methods typically involve detection, feature extraction, and data association steps. SORT [1] uses Kalman filtering for motion prediction and the Hungarian algorithm for data association, achieving real-time performance. DeepSORT [2] extends this approach by incorporating appearance features from deep neural networks to improve identity association. More recent approaches like JDE [3] and FairMOT [4] propose joint detection and embedding frameworks that simultaneously learn object detection and appearance feature extraction in a unified network.

2.2 Multi-Camera Tracking

Multi-camera tracking methods can be broadly categorized into two groups: centralized approaches and decentralized approaches. Centralized approaches [5, 6] first perform tracking in each camera view independently and then associate trajectories across cameras based on appearance and spatio-temporal cues. Decentralized approaches [7, 8] directly establish correspondences between detections from different camera views and perform tracking in a unified space.

Xu et al. [5] proposed a weighted triplet loss to learn discriminative appearance features across different camera views. Chen et al. [6] introduced a cross-view adaptation approach that aligns feature distributions across cameras. Hou et al. [7] developed a multi-view tracker that fuses appearance and geometric features in a unified framework. More recently, Nguyen et al. [8] proposed a graph-based spatio-temporal approach that models both spatial and temporal dependencies for multi-camera tracking.

2.3 Feature Fusion for Tracking

Feature fusion has been widely explored in the tracking literature to combine complementary information from different sources. Early methods used fixed weighting schemes [9] or heuristic rules [10] to combine features. Recent approaches leverage attention mechanisms [11] or adaptive weighting [12] to dynamically adjust feature importance.

Kuo et al. [9] combined appearance and motion features using a fixed weighting scheme. Chen et al. [10] proposed a cascade of features with heuristic rules for feature selection. Zhu et al. [11] introduced a distractor-aware feature fusion approach that uses attention to focus on discriminative features. Lu et al. [12] developed an adaptive feature fusion method that adjusts feature weights based on their reliability.

Our work extends these approaches by introducing a confidence-aware feature fusion mechanism that dynamically adjusts the importance of different features based on their estimated reliability in the current context. Unlike previous methods that rely on predefined rules or fixed attention mechanisms, our approach learns to adapt feature weights based on scene complexity, occlusion levels, and feature quality.

3 Method

3.1 Overview

Our OmniTrack framework consists of four main components: (1) Multi-view detection and feature extraction, (2) Adaptive feature fusion, (3) Confidence-aware data association, and (4) Trajectory refinement with temporal consistency constraints. Figure 1 provides an overview of our approach.

The input to our system is a set of synchronized video streams from multiple cameras with overlapping fields of view. For each frame, we detect people in each camera view using a pre-trained detector and extract appearance, motion, and spatial features. We then project the detections to a common ground plane using camera calibration information and perform feature fusion with our adaptive weighting mechanism. Next, we associate detections across time using our confidence-aware association algorithm. Finally, we refine the trajectories using temporal consistency constraints to improve tracking continuity.

3.2 Multi-View Detection and Feature Extraction

For each camera view $c \in \{1, 2, ..., C\}$ and frame t, we detect people using Faster R-CNN [13] with a ResNet-50 [14] backbone. Each detection $d_{c,t}^i$ consists of a bounding box $b_{c,t}^i = [x, y, w, h]$, a confidence score $s_{c,t}^i$, and the feet position $p_{c,t}^i = [x_f, y_f]$ in the image plane. We project the feet position to the global ground plane using the homography matrix H_c for camera c:

$$P_t^i = H_c \cdot [x_f, y_f, 1]^T \tag{1}$$

where $P_t^i = [X, Y, 1]^T$ represents the position in the global coordinate system.

For each detection, we extract the following features:

1. Appearance features: We use a ResNet-50 model pre-trained on a person re-identification dataset to extract a 2048-dimensional feature vector $f_{c,t,i}^a$ from the detection crop.
2. Motion features: We compute the velocity and acceleration of each detection in the ground plane to form a 4-dimensional motion feature vector $f_{t,i}^m$.
3. Spatial-temporal features: We encode the position and time information in a 3-dimensional vector $f_{t,i}^s = [X, Y, t]$.

To handle the varying reliability of detections from different camera views, we compute a view-specific confidence score $\alpha_{c,t,i}$ for each detection:

$$\alpha_{c,t,i} = s_{c,t}^i \cdot e^{-\lambda \cdot o_{c,t}^i} \tag{2}$$

where $s_{c,t}^i$ is the detection confidence, $o_{c,t}^i$ is the occlusion ratio estimated based on depth ordering, and λ is a hyperparameter controlling the influence of occlusion.

3.3 Adaptive Feature Fusion

Unlike existing methods that use fixed weights for feature fusion, we propose an adaptive fusion mechanism that dynamically adjusts feature weights based on their estimated reliability in the current context. Our fusion approach consists of two steps: (1) Cross-view feature fusion and (2) Multi-feature adaptive fusion.

3.3.1 Cross-View Feature Fusion

For a person detected in multiple camera views, we aggregate the appearance features across views using a confidence-weighted average:

$$\hat{f}_{t,i}^a = \frac{\sum_{c=1}^{C} \alpha_{c,t,i} \cdot f_{c,t,i}^a}{\sum_{c=1}^{C} \alpha_{c,t,i}} \tag{3}$$

where $\hat{f}_{t,i}^a$ is the fused appearance feature for person i at time t.

3.3.2 Multi-Feature Adaptive Fusion

We combine appearance, motion, and spatial-temporal features using an adaptive weighting mechanism:

$$f_{t,i} = w_{t,i}^a \cdot \hat{f}_{t,i}^a + w_{t,i}^m \cdot f_{t,i}^m + w_{t,i}^s \cdot f_{t,i}^s \tag{4}$$

where $w_{t,i}^a$, $w_{t,i}^m$, and $w_{t,i}^s$ are the adaptive weights for appearance, motion, and spatial-temporal features, respectively.

The adaptive weights are computed using a small neural network that takes as input the feature quality indicators and scene complexity metrics:

$$[w_{t,i}^a, w_{t,i}^m, w_{t,i}^s] = \text{softmax}(g_\theta(q_{t,i}^a, q_{t,i}^m, q_{t,i}^s, \eta_t)) \tag{5}$$

where g_θ is a two-layer perceptron with parameters θ, $q_{t,i}^a$, $q_{t,i}^m$, and $q_{t,i}^s$ are quality indicators for each feature type, and η_t is a scene complexity metric that includes crowd density and overall occlusion level.

The feature quality indicators are defined as follows:

$$q_{t,i}^a = \max_{c \in \{1,\ldots,C\}} \alpha_{c,t,i} \tag{6}$$

$$q_{t,i}^m = e^{-\beta \cdot \text{var}(v_{t-k:t,i})} \tag{7}$$

$$q_{t,i}^s = \frac{1}{1 + \gamma \cdot d_{t,i}} \tag{8}$$

where $\text{var}(v_{t-k:t,i})$ is the variance of velocity over the past k frames, $d_{t,i}$ is the distance to the nearest detection, and β and γ are hyperparameters.

3.4 Confidence-Aware Association

We formulate the data association problem as a bipartite matching problem between detections in the current frame and existing trajectories. Our key contribution is the Confidence-Aware Association (CAA) algorithm that explicitly models tracking uncertainty and uses it to guide the association process.

For each trajectory T_j and detection d_i in the current frame, we compute an association cost:

$$C(T_j, d_i) = -\log(p(d_i|T_j)) \tag{9}$$

where $p(d_i|T_j)$ is the probability that detection d_i belongs to trajectory T_j.

We model this probability as a mixture of Gaussians:

$$p(d_i|T_j) = \pi^a \cdot p^a(d_i|T_j) + \pi^m \cdot p^m(d_i|T_j) + \pi^s \cdot p^s(d_i|T_j) \tag{10}$$

where p^a, p^m, and p^s are the probabilities based on appearance, motion, and spatial-temporal features, respectively, and π^a, π^m, and π^s are the corresponding mixture weights.

The appearance probability is computed using the cosine similarity between feature vectors:

$$p^a(d_i|T_j) = \frac{1}{Z^a} \exp\left(\frac{\langle f_{t,i}, f_{t-1,j} \rangle}{\|f_{t,i}\| \cdot \|f_{t-1,j}\| \cdot \sigma^a}\right) \tag{11}$$

where Z^a is a normalization constant and σ^a is a temperature parameter.

The motion probability is computed using a Kalman filter prediction:

$$p^m(d_i|T_j) = \mathcal{N}(P_{t,i}; \hat{P}_{t|t-1,j}, \Sigma_{t|t-1,j}) \tag{12}$$

where $\hat{P}_{t|t-1,j}$ is the predicted position of trajectory T_j at time t and $\Sigma_{t|t-1,j}$ is the corresponding covariance matrix.

The spatial-temporal probability is based on the proximity in the ground plane:

$$p^s(d_i|T_j) = \exp\left(-\frac{\|P_{t,i} - P_{t-1,j}\|^2}{2\sigma^s}\right) \tag{13}$$

where σ^s is a scaling parameter.

The mixture weights π^a, π^m, and π^s are computed as:

$$[\pi^a, \pi^m, \pi^s] = \text{softmax}([\log(q^a_{t,i}), \log(q^m_{t,i}), \log(q^s_{t,i})]) \tag{14}$$

We solve the assignment problem using the Hungarian algorithm [15] with the cost matrix C. To handle new trajectories and false detections, we introduce dummy nodes with a cost threshold τ. If the cost of assigning a detection to any trajectory exceeds τ, the detection initiates a new trajectory.

3.5 Trajectory Refinement with Temporal Consistency

To improve tracking continuity and reduce identity switches, we introduce temporal consistency constraints that exploit the smooth dynamics of human motion. We formulate trajectory refinement as an energy minimization problem:

$$E(T) = \sum_{i=1}^{N} E_{\text{data}}(T_i) + \lambda_1 \sum_{i=1}^{N} E_{\text{smooth}}(T_i) + \lambda_2 \sum_{i \neq j} E_{\text{inter}}(T_i, T_j) \tag{15}$$

where E_{data} is the data term that measures how well the trajectory fits the detections, E_{smooth} is the smoothness term that penalizes non-smooth trajectories, and E_{inter} is the interaction term that penalizes trajectory crossings and overlaps.

The data term is defined as:

$$E_{\text{data}}(T_i) = \sum_{t \in T_i} \|P_{t,i} - \hat{P}_{t,i}\|^2 \tag{16}$$

where $P_{t,i}$ is the observed position at time t and $\hat{P}_{t,i}$ is the refined position.

The smoothness term is defined as:

$$E_{\text{smooth}}(T_i) = \sum_{t \in T_i} \|\hat{P}_{t,i} - 2\hat{P}_{t-1,i} + \hat{P}_{t-2,i}\|^2 \tag{17}$$

This term penalizes acceleration, encouraging smooth trajectories.

The interaction term is defined as:

$$E_{\text{inter}}(T_i, T_j) = \sum_{t \in T_i \cap T_j} \exp\left(-\frac{\|\hat{P}_{t,i} - \hat{P}_{t,j}\|^2}{2\sigma^2}\right) \tag{18}$$

This term penalizes trajectories that are too close to each other, helping to resolve identity switches.

We minimize the energy function using gradient descent with momentum:

$$\hat{P}^{(k+1)}_{t,i} = \hat{P}^{(k)}_{t,i} - \alpha \nabla E(\hat{P}^{(k)}_{t,i}) + \mu(\hat{P}^{(k)}_{t,i} - \hat{P}^{(k-1)}_{t,i}) \tag{19}$$

where α is the learning rate and μ is the momentum coefficient.

4 Experimental Results

4.1 Datasets

We evaluate our approach on three public multi-camera tracking datasets:

1. WILDTRACK [16]: A 7-camera dataset capturing a pedestrian plaza with challenging occlusions and crowded scenes. It contains 400 frames of synchronized videos with calibrated cameras and annotated ground truth.
2. CAMPUS [5]: A 3-camera dataset of a university campus with moderate crowd density. It contains 1500 frames with calibrated cameras and annotated ground truth.
3. PETS2009 [17]: A benchmark dataset for multi-camera tracking with varying crowd densities. We use the S2.L1 scenario with 7 cameras.

4.2 Evaluation Metrics

We use the standard CLEAR MOT metrics [18] for evaluation:

1. Multiple Object Tracking Accuracy (MOTA): A comprehensive metric that combines false positives, false negatives, and identity switches.
2. Multiple Object Tracking Precision (MOTP): The average distance between the predicted positions and the ground truth positions.
3. Identity F1 Score (IDF1): A measure of how well the tracker identifies the correct targets, considering both precision and recall.
4. Mostly Tracked (MT): The percentage of ground truth trajectories that are covered by the tracker for at least 80% of their length.
5. Mostly Lost (ML): The percentage of ground truth trajectories that are covered by the tracker for less than 20% of their length.
6. False Positives (FP): The number of false detections.
7. False Negatives (FN): The number of missed detections.
8. Identity Switches (IDS): The number of times the tracker incorrectly changes the identity of a target.

4.3 Implementation Details

We implement our approach using PyTorch. For person detection, we use Faster R-CNN with a ResNet-50 backbone pre-trained on MS COCO and fine-tuned on each dataset. For appearance feature extraction, we use a ResNet-50 model pre-trained on Market-1501 [19] and DukeMTMC-reID [20].

The adaptive feature fusion network consists of a two-layer perceptron with 64 hidden units and ReLU activation. We train the network using the Adam optimizer with a learning rate of 0.001 and a batch size of 32.

For trajectory refinement, we use gradient descent with a learning rate of 0.05 and a momentum coefficient of 0.9. We set the hyperparameters $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$ based on validation performance.

We perform data augmentation during training, including random horizontal flipping, random cropping, and color jittering. We train our model for 50 epochs on each dataset, with early stopping based on validation performance.

4.4 Results

Table 1 shows the overall performance of our approach compared to state-of-the-art methods on the three datasets.

Our approach achieves the best performance on all three datasets across all metrics. On the WILDTRACK dataset, which features challenging crowd scenarios, OmniTrack achieves a MOTA score of 76.8%, outperforming the previous state-of-the-art method by 2.6%. The improvements are particularly notable in terms of identity switches (IDS), where our method reduces the number by 10.5% compared to GST [8].

On the CAMPUS dataset, OmniTrack achieves a MOTA score of 68.3%, representing a 3.1% improvement over the previous best method. The improvements are consistent across all metrics, with notable reductions in false positives and identity switches.

Table 1: Comparison with state-of-the-art methods on multiple datasets

Dataset	Method	MOTA↑	MOTP↑	IDF1↑	MT↑	ML↓	FP↓	IDS↓
WILDTRACK	DMCT [5]	70.2	75.6	68.9	57.3	14.2	1238	95
	MFMC [6]	73.5	76.2	71.4	60.8	12.7	1173	82
	GST [8]	74.2	77.1	72.6	62.1	11.9	1092	76
	OmniTrack (Ours)	**76.8**	**78.4**	**74.5**	**64.7**	**10.5**	**1021**	**68**
CAMPUS	DMCT [5]	61.4	72.3	59.2	48.6	20.3	1746	143
	MFMC [6]	64.7	73.5	62.1	51.8	18.7	1627	128
	GST [8]	65.2	74.1	63.5	53.6	17.2	1581	119
	OmniTrack (Ours)	**68.3**	**75.8**	**65.7**	**56.2**	**16.1**	**1495**	**105**
PETS2009	DMCT [5]	78.3	79.2	76.4	67.5	10.8	834	62
	MFMC [6]	80.6	80.5	78.2	69.3	9.6	791	54
	GST [8]	81.9	81.3	79.5	71.2	8.7	752	48
	OmniTrack (Ours)	**83.7**	**82.6**	**81.2**	**73.8**	**7.9**	**721**	**41**

On the PETS2009 dataset, our approach achieves a MOTA score of 83.7%, outperforming GST by 1.8%. The relatively smaller improvement on this dataset can be attributed to its less crowded nature, which makes it easier for existing methods to perform well.

Table 2 shows the performance of our approach in different scenarios on the WILDTRACK dataset, categorized by crowd density.

Table 2: Performance in different scenarios on WILDTRACK dataset

Scenario	MOTA↑	MOTP↑	IDF1↑	FP↓	IDS↓
Low density (1-5 people)	85.2	82.5	83.7	187	8
Medium density (6-10 people)	78.6	79.3	76.4	342	21
High density (>10 people)	66.5	73.5	63.4	492	39

As expected, tracking performance decreases as crowd density increases due to more frequent occlusions and interactions. However, our approach maintains reasonable performance even in high-density scenarios, with a MOTA score of 66.5%.

5 Ablation Studies

We conduct ablation studies to analyze the contribution of each component of our approach. Table 3 shows the results on the WILDTRACK dataset.

Table 3: Ablation study on the WILDTRACK dataset

Method	MOTA↑	MOTP↑	IDF1↑	FP↓	IDS↓
OmniTrack (Full)	76.8	78.4	74.5	1021	68
- Adaptive fusion	74.3	77.2	72.1	1109	81
- Confidence-aware association	73.7	76.8	71.5	1132	87
- Temporal consistency	75.2	77.9	73.2	1054	79
- All components (baseline)	71.8	75.3	69.4	1196	93

Removing the adaptive fusion mechanism results in a 2.5% decrease in MOTA, highlighting the importance of dynamically adjusting feature weights based on scene conditions. When the confidence-aware association is replaced with standard association, MOTA decreases by 3.1%, with a notable increase in identity switches. This confirms the effectiveness of our approach in handling uncertainty during data association.

Removing the temporal consistency constraints reduces MOTA by 1.6%, with a smaller impact compared to the other components. This suggests that while temporal consistency is beneficial, the adaptive fusion and confidence-aware association provide more substantial improvements.

When all components are removed (resulting in our baseline implementation), MOTA decreases by 5.0%, demonstrating the combined effect of our contributions.

We further analyze the adaptive feature fusion mechanism by examining the average weights assigned to different features across various scenarios. Table 5 shows the results.

Table 4: Average feature weights in different scenarios

Scenario	Appearance	Motion	Spatial-temporal
Low occlusion	0.58	0.27	0.15
Medium occlusion	0.42	0.38	0.20
High occlusion	0.31	0.46	0.23
Low crowd density	0.53	0.30	0.17
Medium crowd density	0.44	0.36	0.20
High crowd density	0.32	0.43	0.25

The results in Table 5 reveal important insights about our adaptive feature fusion mechanism. In scenarios with low occlusion, the appearance features receive the highest weight (0.58), as they are most reliable when people are clearly visible. As occlusion increases, the weight shifts toward motion features (0.46 in high occlusion scenarios) and spatial-temporal features (0.23 in high occlusion scenarios), which are more robust under occlusion.

Similarly, in low crowd density scenarios, appearance features dominate (0.53), but as crowd density increases, the weights shift toward motion and spatial-temporal features. This adaptive weighting allows our approach to maintain robust performance across diverse scenarios by emphasizing the most reliable features in each context.

We also analyze the impact of the number of cameras on tracking performance. Figure 2 shows the MOTA scores with varying numbers of cameras on the WILDTRACK dataset.

As expected, tracking performance improves with more cameras due to better coverage and reduced occlusions. Our approach shows a steeper improvement curve compared to baseline methods, demonstrating its effectiveness in leveraging information from multiple views. With all seven cameras, OmniTrack achieves a MOTA score of 76.8%, while the performance drops to 66.2% with only three cameras.

To understand the effectiveness of our Confidence-Aware Association (CAA) algorithm, we analyze its performance under different crowd densities. Table 6 shows the number of identity switches per 100 frames for different association methods.

Table 5: Average feature weights in different scenarios

Scenario	Appearance	Motion	Spatial-temporal
Low occlusion	0.58	0.27	0.15
Medium occlusion	0.42	0.38	0.20
High occlusion	0.31	0.46	0.23
Low crowd density	0.53	0.30	0.17
Medium crowd density	0.44	0.36	0.20
High crowd density	0.32	0.43	0.25

The results in Table 5 reveal important insights about our adaptive feature fusion mechanism. In scenarios with low occlusion, the appearance features receive the highest weight (0.58), as they are most reliable when people are clearly visible. As occlusion increases, the weight shifts toward motion features (0.46 in high occlusion scenarios) and spatial-temporal features (0.23 in high occlusion scenarios), which are more robust under occlusion.

Similarly, in low crowd density scenarios, appearance features dominate (0.53), but as crowd density increases, the weights shift toward motion and spatial-temporal features. This adaptive weighting allows our approach to maintain robust performance across diverse scenarios by emphasizing the most reliable features in each context.

We also analyze the impact of the number of cameras on tracking performance. Figure 2 shows the MOTA scores with varying numbers of cameras on the WILDTRACK dataset.

As expected, tracking performance improves with more cameras due to better coverage and reduced occlusions. Our approach shows a steeper improvement curve compared to baseline methods, demonstrating its effectiveness in

leveraging information from multiple views. With all seven cameras, OmniTrack achieves a MOTA score of 76.8%, while the performance drops to 66.2% with only three cameras.

To understand the effectiveness of our Confidence-Aware Association (CAA) algorithm, we analyze its performance under different crowd densities. Table 6 shows the number of identity switches per 100 frames for different association methods.

Table 6: Identity switches per 100 frames for different association methods under varying crowd densities

Method	Low density	Medium density	High density
Hungarian with fixed weights	2.3	5.1	10.9
Hungarian with adaptive weights	1.8	4.3	8.6
Confidence-Aware Association (Ours)	**1.1**	**3.2**	**6.1**

The results demonstrate that our CAA algorithm significantly reduces identity switches, especially in high-density scenarios. Compared to the standard Hungarian algorithm with fixed weights, CAA reduces identity switches by 52.2% in low-density scenarios and 44.0% in high-density scenarios. This improvement can be attributed to the explicit modeling of uncertainty in the association process and the adaptive mixture weights for different feature types.

6 Discussion

The experimental results highlight several key strengths of our approach. First, the adaptive feature fusion mechanism effectively adjusts to varying scene conditions, maintaining robust performance across different crowd densities and occlusion levels. This adaptivity is particularly valuable in real-world scenarios where conditions can change rapidly.

Second, the Confidence-Aware Association algorithm significantly reduces identity switches by explicitly modeling uncertainty and using it to guide the data association process. This is crucial for maintaining consistent trajectories in crowded environments where standard association methods often fail.

Third, the temporal consistency constraints provide additional refinement that improves tracking continuity, reducing fragmented trajectories and further enhancing overall performance.

Despite these strengths, our approach has several limitations that warrant further investigation. First, the current implementation relies on pre-trained detectors and feature extractors, which may not be optimized for the specific characteristics of each camera view. End-to-end training of the entire pipeline could potentially improve performance but would require more extensive computational resources.

Second, while our method adapts to different crowd densities, extreme crowding still poses significant challenges, as indicated by the performance drop in high-density scenarios. Incorporating additional cues such as human pose or depth information could potentially improve robustness in these cases.

Third, our approach assumes calibrated cameras with overlapping fields of view. Extending the method to handle uncalibrated cameras or limited overlap would broaden its applicability to more general surveillance scenarios.

7 Conclusion

In this paper, we presented OmniTrack, a novel framework for multi-camera human tracking in crowded environments. Our approach leverages adaptive feature fusion and temporal consistency constraints to improve tracking performance, particularly in challenging scenarios with occlusions and similar appearances. The key innovations include an adaptive feature fusion mechanism that dynamically adjusts feature weights based on their reliability, a Confidence-Aware Association algorithm that explicitly models tracking uncertainty, and temporal consistency constraints that exploit the smooth dynamics of human motion.

Extensive experiments on three public datasets demonstrate that our approach achieves comparable or superior performance to state-of-the-art methods, with notable improvements in crowded scenes with frequent occlusions. The ablation studies confirm the contribution of each component to the overall performance and provide insights into their effectiveness in different scenarios.

Future work will explore several directions to address the limitations discussed above. First, we plan to investigate end-to-end training of the entire pipeline to optimize detection and feature extraction for specific camera views. Second, we will explore the integration of additional cues such as human pose and depth information to improve robustness in

extremely crowded scenes. Third, we will extend our approach to handle uncalibrated cameras and limited overlap to broaden its applicability.

In summary, OmniTrack provides a robust solution for multi-camera human tracking in crowded environments, with adaptive features that make it suitable for a wide range of real-world applications. The demonstrated improvements over existing methods highlight the effectiveness of our approach and its potential impact on surveillance, sports analytics, and other domains that rely on accurate human tracking.

8 References

References

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464-3468.

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

[3] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 107-122.

[4] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," in *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069-3087, 2021.

[5] Y. Xu, X. Liu, Y. Liu, and S. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4256-4265.

[6] X. Chen, K. Xiang, X. Song, and Q. Huang, "Cross-view tracking for multi-human 3D pose estimation at over 100 FPS," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3276-3285.

[7] Z. Hou, X. Wu, J. Sun, and H. Jia, "Multiview detection with feature perspective transformation," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 1-18.

[8] N. Nguyen, S. Lan, and R. Wang, "Graph-based spatio-temporal multi-camera tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11534-11543.

[9] C. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 685-692.

[10] C. Chen, A. Heili, and J. Odobez, "A multi-feature framework with adaptive cascading weights for long-term visual tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2102-2106.

[11] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 101-117.

[12] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M. H. Yang, "Adaptive online correlation tracking with dual memory," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1666-1680, 2020.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91-99.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[15] H. W. Kuhn, "The Hungarian method for the assignment problem," in *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83-97, 1955.

[16] T. Chavdarova et al., "WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5030-5039.

[17] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, 2009, pp. 1-6.

[18] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," in *Journal on Image and Video Processing*, vol. 2008, pp. 1-10, 2008.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116-1124.

[20] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision Workshops (ECCVW)*, 2016, pp. 17-35.

[21] L. Leal-Taixé et al., "MOTChallenge 2015: Towards a benchmark for multi-target tracking," in *arXiv preprint arXiv:1504.01942*, 2015.

[22] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," in *arXiv preprint arXiv:1603.00831*, 2016.

[23] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3D pose estimation from multiple views," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7792-7801.

[24] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 854-865, 2019.

[25] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1-8.

[26] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," in *arXiv preprint arXiv:2107.08430*, 2021.

[27] R. E. Kalman, "A new approach to linear filtering and prediction problems," in *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35-45, 1960.

[28] J. Wang, Y. Jiang, Z. Wang, Z. Zhang, Y. Zhao, and Z. Chen, "Exploit the connectivity: Multi-object tracking with trackletnet," in *ACM International Conference on Multimedia*, 2019, pp. 482-490.