OCMCTrack: Online Multi-Target Multi-Camera Tracking with Corrective Matching Cascade

Andreas Specker
Fraunhofer IOSB
Fraunhofer Center for Machine Learning
{andreas.specker@iosb.fraunhofer.de}

Abstract

The implementation of multi-target multi-camera tracking systems in indoor environments, including shops and warehouses, facilitates strategic product positioning and the improvement of operational workflows. This paper presents the online multi-target multi-camera tracking framework OCMCTrack, which tracks the 3D positions of people in the world. The proposed framework introduces a novel matching cascade to re-evaluate track assignments dynamically, thus minimizing false positive associations often made by online trackers. Additionally, this work presents three effective methods to enhance the transformation of a person's position in the image to world coordinates, thereby addressing common inaccuracies in positional reference points. The proposed methodology is able to achieve competitive performance in Track 1 of the 2024 AI City Challenge, demonstrating the effectiveness of the framework.

1. Introduction

The objective of multi-target multi-camera tracking (MTMCT) is to determine the trajectories of multiple entities, such as people, vehicles, or other objects, within an environment monitored by an array of cameras. The growing interest in MTMCT in indoor environments [15] is driven by its potential to significantly impact various domains, including security, retail, industrial, and logistic sectors. The ability to track multiple entities, such as humans or objects, across several camera views allows for a comprehensive understanding of movement patterns and interactions within the monitored area. In retail, MTMCT systems facilitate the analysis of consumer behavior, enabling the optimization of shop layouts and product placement. Within industry, MTMCT can improve logistic operations, streamline manufacturing, and enhance the safety of workers.

The task of MTMCT can be conceptually divided into two main components: tracking within single-camera views, also referred to as multi-object tracking (MOT), and cross-camera association, as visualized in Fig. 1. Tracking within a camera view involves the detection of individuals within the camera's field of view, and the subsequent linking of these detections across successive frames to create single-camera tracks. Afterward, cross-camera association merges these single-camera tracks captured by different cameras to construct multi-camera trajectories that encompass an individual's movement across the entire array of cameras.

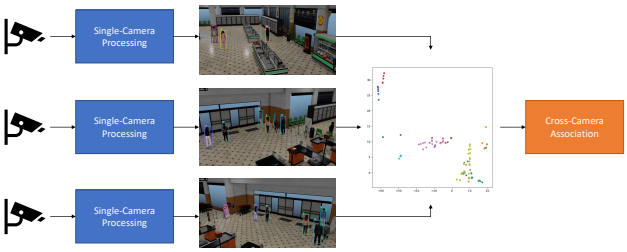


Figure 1. **Framework Overview** – First, single-camera processing components generate tracks within camera views separately for each camera. These single-camera tracks are then projected into world coordinates. Cross-camera association is performed in world coordinates to allow single-camera processing on the edge and flexible integration of new cameras without changes to the multi-camera tracker.

This work aims to accurately locate subjects within a global world coordinate system, rather than tracking individuals in local image coordinate systems. This introduces additional complexities to the already challenging MTMCT task. Varying lighting conditions, occlusions, and the complex nature of human motion are classical challenges in MTMCT. These challenges are extended by the need of accurate transformation of person detections into the world coordinate system.

The proposed online MTMCT system OCMCTrack, which stands for online corrective matching cascade track-

ing, is designed with modularity, flexibility, and scalability in mind to ensure the system can adapt to various real-world environments and to allow upgrades by recent research advancements. The framework separates image-based processing components from cross-camera association entirely, prioritizing privacy as a major concern when deploying MTMCT systems. As shown in Fig. 1, cross-camera association is executed exclusively on non-image data, such as global coordinates and visual appearance features, allowing image-based processing on the edge. Thus, storing or transferring image data through the network is not required, embedding privacy-by-design into the system architecture. Furthermore, network communication is limited to lightweight positional information and feature vectors, which is expected to significantly reduce latency.

Online tracking methods have a major drawback compared to offline methods: they might lack sufficient information when making across-camera association decisions. To close the gap between the accuracy of online and offline trackers, this work proposes to re-evaluate the assignment of single-camera tracks to multi-camera tracks in each time step, thus, reducing false positive associations. Furthermore, this work presents three effective methods for improving the accuracy of bounding box to world position transformations and compensating for inaccurate reference points during transformation.

The primary contributions of this work are summarized as follows:

- The introduction of a modular, flexible, and efficient online MTMCT framework that is designed to serve as a baseline for subsequent research
- A novel matching cascade that incorporates a mechanism to correct erroneous associations from previous time steps
- Three efficient approaches aimed at resolving the challenge of determining plausible world coordinates from bounding boxes within video frames
- Competitive results in Track 1 of the 2024 AI City Challenge [27]

2. Related Work

In the field of MTMCT, modern methods adopt a two-staged approach: single-camera tracking to monitor the routes of individuals within a single camera's field of view, followed by inter-camera tracking to match the resulting tracklets across the camera network [7, 8, 11, 13, 19, 22, 24, 26, 30]. The two-stage procedure has become a widely accepted paradigm due to its effectiveness in disentangling the complexity of MTMCT. The first stage, single-camera tracking, has been extensively studied, with the tracking-by-detection paradigm prevailing as the predominant focus [1, 4, 10, 28, 34]. This approach detects individuals in each camera frame first and then links the detections over time to maintain their identity across frames.

Similar to single-camera tracking algorithms, multi-camera tracking approaches can be broadly classified into online [6,20,21,33] and offline methodologies [7,11,13,22, 24,30]. Online methods [6,20,33] address the tracking challenge in a sequential frame-by-frame manner, thereby facilitating the potential for real-time processing and immediate tracking results, which is highly desirable in real-world applications. In contrast, offline approaches use the entirety of data output from single-camera tracking algorithms to find the best association across cameras [7,11,13,22,24,30]. Offline methods are preferred in competitive contexts, such as the AI City Challenge [15,16], due to their higher accuracy. This is because they can leverage comprehensive temporal information to resolve ambiguities in track associations.

Recent developments in the field emphasize the crucial role of scene-specific prior knowledge in MTMCT [7,8,11, 18, 20, 22, 24]. Knowing camera positions and orientations is essential to prevent unrealistic track associations between camera views, particularly when dealing with overlapping fields of view or when aiming at determining the position of objects in a global world coordinate system. A number of studies have also proposed the use of spatial projections onto a ground plane as a method to enhance tracking precision [3, 17, 25, 29].

The approach presented in this research aligns with contemporary findings. It adopts a two-stage online strategy and the tracking-by-detection scheme. Furthermore, geometric transformations are leveragedto project bounding boxes to the ground plane to get global positions in the world coordinate system.

3. Methodology

The fundamental principle of the proposed online multi-camera tracking framework is shown in Fig. 1. It comprises three main processing steps: single-camera processing, bounding box projection, and cross-camera association. This design is chosen due to multiple reasons. First, this allows computation of video streams on the edge, which has several advantages. There is no need to transfer large amounts of video data to the centralized cross-camera association module but only lightweight global positions with corresponding visual feature vectors instead. As a result, latency for online tracking is greatly reduced. Furthermore, this procedure is more privacy-friendly as video data showing persons is only available on the edge and is deleted directly after the processing stopped. Last, the proposed framework offers flexibility and modularity as new cameras can be added without any changes to the cross-camera

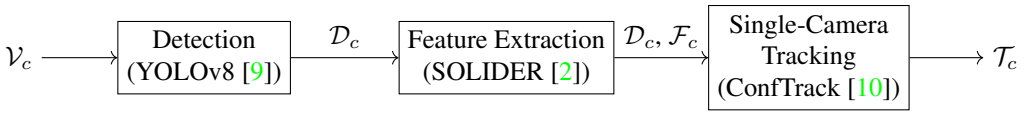


Figure 2. **Single-camera Tracking** – First, a detector localizes the persons in the frames. The feature extractor generates visual embeddings of the detected persons' appearances. Finally, the single-camera tracker ConfTrack [10] is applied to associate the detections over time to produce single-camera tracks.

component even during runtime. Similarly, single-camera processing modules can be updated without effects to the cross-camera tracking.

The following sections elaborate on the three main components in detail.

3.1. Single-Camera Processing

As depicted in Fig. 2, the single-camera processing module receives a video stream $\mathcal{V}_c$ captured by the c-th camera as input and produces a set of single-camera tracks $\mathcal{T}_c$. The processing pipeline follows the tracking-by-detection paradigm and consists of separate detection, appearance feature extraction, and single-camera tracking stages. The detection stage localizes persons in the video frames and creates a set of bounding boxes $\mathcal{D}_c$ that are enriched with appearance feature vectors $\mathcal{F}_c$ extracted using a person re-identification model. This information is provided as input to an online single-camera tracking method to produce the final single-camera tracks.

3.1.1 Detection

Due to its favorable trade-off between computation time, accuracy, and generalization ability [5, 21, 23], the single-stage YOLOv8x detection model [9] was selected as the default person detector in OCMCTrack. The model is initialized with COCO pre-trained weights and fine-tuned using the adapted bounding boxes described in Sec. 3.2. The input videos are processed in their original resolution, *i.e.* 1920×1080 pixels. Except for the learning rate, which is reduced to 0.001, the default parameters give the best performance.

3.1.2 Feature extraction

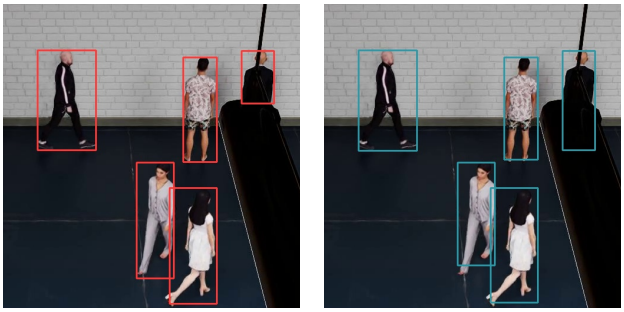
The feature extraction step computes visual embeddings $\mathcal{F}_c$ for bounding boxes included in $\mathcal{D}_c$. Typically, person re-identification models are applied for this task that are trained to learn a feature space in which embeddings of the same individual seen from different views are close while embeddings of distinct persons are far away. Recent advancements in computer vision demonstrated great potential of unsupervised pre-training and transformer architectures. Therefore, SOLIDER [2] serves as appearance feature extractor in this work. The model uses the Swin [12] transformer as backbone and is pre-trained in an unsupervised manner using diverse imagery from the internet. Fine-tuning is performed on the AI City Challenge 2024 dataset with the original parameter setting, except for the batch size (128) and the number of individuals within a batch (8). The *Base* variant of the Swin transformer is utilized as the backbone model.

3.1.3 Single-Camera Tracking

Many tracking-by-detection algorithms are proposed recently that add incremental improvements to their predecessors. Thus, one of the most current methods, concretely ConfTrack [10], builds the basis for the single-camera tracking stage in the proposed framework. ConfTrack contributes multiple advancements, out of which only three proved beneficial. On the one hand, the noise scale adaptive Kalman filter update amplifies the measurement noise of the detected box dependent on the predicted confidence score. As a result, the Kalman update focuses more on the predicted bounding box than on noisy ones from the detector and the tracking accuracy improves. Moreover, the authors propose to predict the size of bounding boxes constantly after a track is lost. Furthermore, ConfTrack also exploits the confidence scores from the detector to adjust the matching cost. The lower the confidence of a detected bounding box, the higher the cost for matching with an existing track. This procedure prefers matches with certainly recognized boxes and avoids false positive assignments. However, the confidence-weighted Kalman update (CWKU) and the additional low-confidence stage in the matching cascade (LCTM) are removed based on empirical observations.

3.2. Projection

The goal of the challenge is to determine global positions of persons in the scene. To achieve this, person detections in the camera views need to be projected into the world coordinate system. Given the homography matrix $\mathbf{H}_c$ for camera c, the 2D world position $[x, y]^T$ of the image coordinates



(a) Original boxes (b) Adjusted boxes

Figure 3. **Adjustment of Bounding Boxes** – The original bounding box annotations (Fig. 3a) are adapted (Fig. 3b) to improve the accuracy of bounding box transformation into world coordinates.

$[u, v]^T$ are computed as follows:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = \mathbf{H}_c^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} \div \hat{z} \tag{2}$$

However, this requires the selection of an appropriate reference point for each person detection in a video frame. The straightforward approach is to use the horizontally centered point at the bottom of a bounding box. While this results in high accuracy for standing persons that are entirely visible, localization accuracy drops when the lower-body is occluded and thus not detected or when persons walk towards or away from the camera. Possible solutions include leveraging pose estimation to exactly determine the point of contact with the floor, or extend the detector to additionally predict this position. However, these methods introduce further computational complexity, which harms efficiency and therefore real-time capability. Moreover, modularity suffers since state-of-the-art algorithms may not be applied off-the-shelf and specific models need to be developed.

Based on these considerations, a lightweight yet accurate approach is followed. The training bounding boxes are simply adjusted to span not only over the visible parts of the person, but to have their bottom where the person is located in the world coordinate system. This point is determined using the projection matrix and the global positions of persons provided in the dataset annotations. Original and adjusted bounding boxes are compared in Fig. 3. As can be seen for the bald man in the upper right of Fig. 3b, adapted training bounding box include the entire body even if parts are invisible. In contrast, the original annotation, visualized in Fig. 3a, only spans over the visible part.

However, one error case for inaccurate world position estimates is still not fixed by this. Some detectors such as YOLOv8 are restricted to predicting bounding boxes within the image boundaries. As a result, reference points of partly visible persons at the bottom of frames are not reliable. Therefore, such cases are handled separately by expanding border bounding boxes to determine the reference points for projection. Concretely, bounding boxes are extended to have 2.5 times the height as the width. The enlargement factor correlates with the mean ratio of annotated bounding boxes in the training and validation set of the AI City Challenge 2024 dataset.

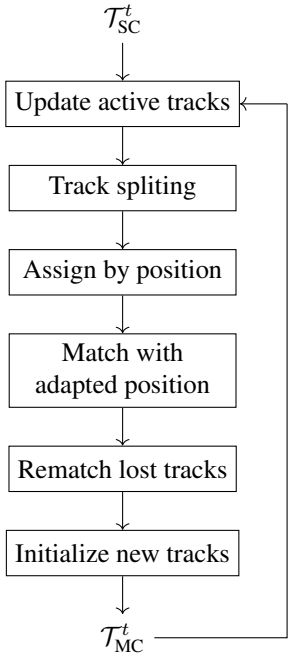


Figure 4. **Cross-camera Association** – Overview of the cross-camera association processing steps.

3.3. Cross-Camera Association

The cross-camera association module links the single-camera tracks based on the projected world positions. Inputs to the processing pipeline are current single-camera tracks $\mathcal{T}_{\text{SC}}^t$ in time step t and multi-camera tracks $\mathcal{T}_{\text{MC}}^{t-1}$ from the previous time step $t-1$. The proposed multi-camera tracker distinguishes between two track states: *active* and *lost*. *Active* multi-camera tracks are currently visible in at least on camera view, while *lost* tracks belong to individuals that are currently out of the cameras' fields of view. The two sets of tracks are forward through the processing pipeline provided in Fig. 4. First, multi-camera tracks included in $\mathcal{T}_{\text{MC}}^{t-1}$ are updated using the new positions of the associated single-camera tracks. Subsequently, single-camera tracks that do not sufficiently match the assigned multi-camera track after the update are disassociated and considered matching candidates to other multi-camera

tracks in the following processing stages. Afterward, previously unmatched single-camera tracks and existing multi-camera tracks are linked by position in two different manners. Next, *lost* multi-camera tracks are rematched. Finally, remaining unmatched single-camera tracks are spatially clustered to form new multi-camera tracks. The result is the new set of multi-camera tracks $\mathcal{T}_{\text{MC}}^{t}$. The remainder of this section provides details about the separate stages.

Update active multi-camera tracks. Existing multi-camera tracks are updated by extracting the world positions of associated single-camera tracks. The current track position is set to the median position of the single-camera tracks. Similarly, current appearance features are retrieved and updated. If all linked single-camera tracks are finished, multi-camera tracks switch to the *lost* state.

Track splitting. The major drawback of online trackers is that association decisions must be made with limited information available. As a result, the risk of false positive or missed links is increased. To alleviate their impact on the tracking results, the track splitting stage examines the single-camera tracks included in a multi-camera track concerning their similarity. Similarity is measured by position and visual appearance. If respective thresholds τ_{p} or τ_{v} are exceeded, outlier single-camera tracks are removed from the multi-camera track to correct faulty association decisions from previous time steps. Rejected single-camera tracks are eligible to be matched with other multi-camera tracks in subsequent stages.

Assign by position. This stage assigns single-camera tracks to multi-camera tracks based on their positions in the world. Unmatched single-camera tracks' and active multi-camera tracks' positions are clustered in a hierarchical manner until the distance criterion τ_{d} is met. As an additional constraint, it is enforced that only one active single-camera track per camera can be clustered, since the same individual cannot appear at multiple locations in the same video frame.

Match with adapted position. Analogous to the previous processing step, association is performed based on the global positions. But in contrast, the positions of single-camera tracks are adjusted to compensate for inaccurate reference points for projection into world coordinates. In detail, a straight line of world positions is computed which corresponds to increasing the height of a bounding box. Then, the shortest distances of the multi-camera tracks' positions to this straight line, *i.e*. possibly more accurate world positions are computed to perform the association. The employed distance threshold is referred to as τ_{a}.

Rematch lost tracks. Unlike the previous stages, this processing stage considers *lost* multi-camera tracks. The goal is to recover the tracks of persons who re-enter the captured scene and camera views after having been invisible for some time. Since the movements and thus the positions of persons in such cases are hardly predictable, the visual appearance is leveraged as the main cue for matching. The distance between the visual embeddings of single- and multi-camera tracks are calculated and these are matched if the distances are below the threshold τ_{r}. Additionally, a time-dependent constraint is applied. The shorter the multi-camera track was invisible, the closer the last position of the multi-camera track and the position of the matching candidate must be. Rematched multi-camera tracks change their state to *active*.

Initialize new tracks. Single-camera tracks that were not associated with a multi-camera track during the previous stages are utilized to initialize new multi-camera tracks. To do this, world positions are clustered hierarchically with the distance threshold τ_{n}. Each resulting cluster forms a new multi-camera tracks. Analogous to the assign by position procedure, only one active single-camera track per camera is allowed within a multi-camera track.

4. Experiments

This section describes the experimental framework and presents the empirical findings. First, we describe the dataset used in this study. Then, we present the hyperparameters and provide a comprehensive analysis of the obtained results.

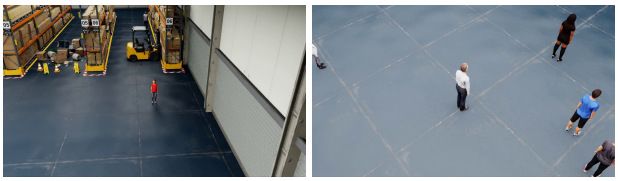
4.1. Dataset

The dataset provided for Track 1 of the 2024 AI City Challenge [27] constitutes a comprehensive, synthetic multi-target multi-camera tracking dataset, generated via the NVIDIA Omniverse Platform. The dataset includes a total of 90 distinct indoor environments, each monitored by multiple strategically positioned cameras. These cameras capture both overlapping and isolated fields of view, providing a diverse range of perspectives.

The video data has a Full HD resolution with a capture rate of 30 frames per second. Samples from various scenarios within the dataset are presented in Figure Fig. 5. The dataset is partitioned into separate subsets for training (40 scenarios), validation (initially 20, reduced to 12 scenarios), and testing (30 scenarios). To increase the amount of training data, scenes originally designated for validation, specifically scenes 49 to 60, are reassigned to the training subset.

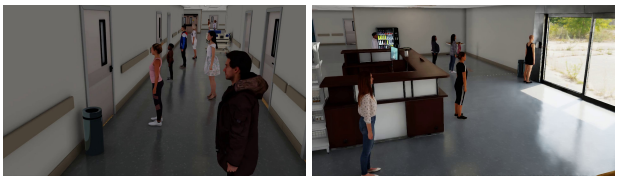
For the development and evaluation of detection algorithms, a subset of the frames, specifically every 60th frame, is utilized. In the context of person re-identification, unique

(a) Scene 1



(b) Scene 61

(c) Scene 71



(d) Scene 81

Figure 5. **AI City Challenge 2024 Dataset** – Selected camera views of four multi-camera scenes of the challenge dataset are shown. The dataset comprises various synthetic scenarios.

τ_p	τ_v	τ_d	τ_a	τ_r	τ_n
1.8	0.1	1.4	1.5	0.18	1.3

Table 1. **Hyperparameter –** Overview of hyperparameter values chosen for the proposed online tracking framework.

identities are sparse relative to their frequency of appearance, so individuals are extracted solely from every 128th frame to maintain data diversity. A single bounding box per individual and video is sampled to serve as a query. Multi-camera tracking experiments are evaluated on scenes 49 to 51.

4.2. Hyperparameters

To mimic a real-world scenario in which it is hardly possible to choose camera-specific parameters for large numbers of cameras and scenarios, a unique set of hyperparameters is selected based on the validation results. The hyperparameter values are summarized in Tab. 1.

Model	Size	mAP50	mAP50-95
DINO [32] Swin-Tiny	1920px	98.6	90.4
DINO [32] Swin-Base	1920px	98.8	92.2
YOLOv8x [9]	1280px	**99.3**	96.6
YOLOv8x [9]	1920px	**99.3**	**97.1**

Table 2. **Detection Results –** Comparison of YOLOv8x models trained with difference image resolutions and DINO models.

4.3. Results & Discussion

This section presents and discusses experimental findings and thorough ablation studies to justify the proposed methodology. Besides, it provides the results of track 1 of the 2024 AI City Challenge. Detection and person re-identification are evaluated based on well-established metrics, while multi-camera tracking is assessed using the official challenge metric. It is an adapted version of the Higher Order Tracking Accuracy (HOTA) metric [14] for the multi-camera setup and 2D global positions. This metric can be decomposed into three subparts: detection accuracy (DetA), association accuracy (AssA), and localization accuracy (LocA)

Detection. The results of the detection algorithms are systematically presented in Tab. 2. It is worth noting that the YOLO architecture clearly outperformed the DINO models, which was unexpected. This could be due to the extensive data augmentation applied during the YOLO training process, which may have helped to prevent overfitting. Further analysis shows a positive correlation between image resolution and detection accuracy, especially in terms of mean Average Precision (mAP) across a range of Intersection over Union (IoU) thresholds from 0.50 to 0.95. This indicates that higher image resolution significantly improves the accuracy of bounding box localization. Given the importance of precise bounding box localization for subsequent transformations into world coordinate space, the YOLOv8x model, which uses Full HD resolution images as input, has been determined to be the optimal standard configuration for the tracking framework.

Person re-identification. A comparative analysis was conducted between the established AGW baseline and the novel SOLIDER methodology concerning person re-identification. The experimental results consistently indicate that all configurations of SOLIDER outperform the AGW baseline across the metrics. Notably, the use of a larger Swin-Base backbone model as part of the SOLIDER framework is associated with significant improvements in both Rank-1 Accuracy (R-1) and Mean Average Precision

Approach	Backbone	R-1	mAP
AGW [31]	ResNet-50	65.2	54.1
AGW [31]	ConvNeXt-Base	65.0	52.7
SOLIDER [2]	Swin-Tiny	67.5	61.4
SOLIDER [2]	Swin-Base	**70.0**	63.0
SOLIDER [2] + BS128i8	Swin-Base	**70.0**	**65.6**

Table 3. **Person Re-identification Results –** Person re-id results achieved on the validation split of the AI City Challenge dataset.

Approach	HOTA	DetA	AssA	LocA
ByteTrack [34]	89.4	90.8	87.9	94.1
ConfTrack [10]	89.9	90.9	88.9	94.1
Ours (adapted ConfTrack [10])	**90.8**	91.0	**90.6**	**94.2**
w/ CWKU	90.5	**91.1**	89.9	**94.2**
w/ LCTM	89.9	90.9	89.0	94.1

Table 4. **Single-Camera Tracking Ablation Study –** Multi-camera tracking results for different variants of single-camera tracking algorithms. Our approach uses ConfTrack without CWKU and LCTM.

(mAP). These metrics indicate the model's ability to accurately identify individuals on the first attempt and its precision across all attempts. Additionally, optimizing the training process by increasing the batch size from 64 to 128 and doubling the number of unique identities sampled per batch from 4 to 8 has resulted in further performance improvements (BS128i8). The results suggest that the SOLIDER approach benefits from larger batch sizes and a greater diversity of individuals within each batch. This likely contributes to a more robust feature learning and generalization capability of the model. Based on this empirical evidence, the SOLIDER approach with a Swin-Base backbone and optimized batch sampling parameters is selected as a superior configuration for person re-identification tasks within the scope of the tracking framework.

Single-Camera Tracking. Tab. 4 evaluates and compares the performance of different single-camera tracking methods. The analyzed data shows that ConfTrack [10] outperforms ByteTrack [34] in terms of tracking accuracy. Comparing ConfTrack algorithm variants reveals that the original ConfTrack algorithm, which includes CWKU and LCTM, is less effective than its adapted version that excludes these components. The experiments confirm that integrating either CWKU or LCTM independently results in a degradation in tracking performance. Therefore, to achieve optimal tracking results, both CWKU and LCTM are excluded from the ConfTrack algorithm in this work.

Approach	HOTA	DetA	AssA	LocA
Ours	**90.8**	91.0	**90.6**	**94.2**
w/o adjusted detection boxes	84.6	86.3	83.1	91.0
w/o enlarged border boxes	87.8	88.8	86.7	93.8
w/o match with adapted position	90.1	90.7	89.5	**94.2**
w/o track splitting	45.7	80.9	26.5	93.8
w/o track splitting (visual)	49.4	85.9	29.0	94.0
w/o track splitting (position)	90.7	**91.1**	90.2	**94.2**

Table 5. **Multi-camera Tracking Ablation Study –** The proposed components for projection and multi-camera tracking are evaluated concerning their influence on multi-camera tracking performance.

Cross-Camera Association. An evaluation of the proposed modules and their impact on multi-camera tracking accuracy is presented in Tab. 5. The first block of ablation results examines the influence of the developed methods for deriving accurate world coordinates from bounding boxes. A notable observation is that using the original bounding box annotations for detector training, as opposed to the adjusted ones, results in a remarkable decrease in all performance metrics. A similar observation is made when the strategy of enlarging bounding boxes at image boundaries is excluded. The least pronounced impact on accuracy metrics is associated with omitting the match with adapted position stage from the matching cascade. The only modest decrease can be attributed to the fact that this stage only affects a few cases. This is mainly due to the effectiveness of the two aforementioned measurements, which greatly mitigate the problem of inaccurate reference points for position transformation.

The experimental analysis focused on track splitting shows that excluding it completely leads to a significant reduction in HOTA. This is predominantly caused by a strong decrease in association accuracy AssA, which drops from 90.6% to only 26.5%. The finding highlights the importance of the track splitting algorithm in maintaining high tracking accuracy. Incorporating this mechanism effectively reduces the impact of false positive associations, making them acceptable within certain limits. As a result, more links can be allowed by loosening the matching thresholds to increase the association recall without negative influence by also increased false positives. To delve deeper into track splitting, results for omitting each of the two splitting criteria individually are investigated. The findings clearly demonstrate that the use of visual information for track splitting is essential for accurate tracking. When only position-based splitting was used, the AssA and HOTA scores dropped significantly to 29.0% and 49.4%, respectively. Excluding positional splitting had a negligible effect on the results. The HOTA metric only experienced a marginal decline of 0.1 percentage points. Visual features tend to become more robust over

Rank	Team	HOTA	Rank	Team	HOTA
1	RIIPS	71.9	9	Asilla	40.3
2	SJTU-Lenovo	67.2	10	TryThis	33.5
3	NetsPresso	60.9	11	Graph@FIT+Comenius	31.5
4	**Ours**	**60.9**	12	Deeper	27.7
5	UWIPL-ETRI	57.1	13	JRZ Vision2Move	23.4
6	ARV RETERIU	51.1	14	lab511	13.2
7	SKKU-AutoLab	45.2	15	SCU_Anastasiu	6.6
8	STCHD	40.6	16	Tahakom	5.2

Table 6. **Challenge results –** Challenge results on the official test set. Online methods, such as the proposed framework, will receive a bonus of 10 percentage points for the final ranking. These points are not included.

time, making them a valuable indicator over time whether a correct or false positive matching was performed. Contrary to that, global position data is highly reliable at initial stage. Therefore, fewer errors occur and splitting based on this criterion is less important.

Challenge Results. Tab. 6 presents the results of the challenge [27]. It is important to note that the scores displayed do not account for the 10 percentage point bonus rewarded to online tracking methods, which includes the proposed framework. The methodology introduced in this work has accomplished a HOTA score of 60.9%. This score has positioned the framework at the fourth rank within Track 1 of the 2024 AI City Challenge.

5. Conclusion

In conclusion, this work presents a novel online MTMCT framework, named OCMCTrack. The framework's design is intended to serve as a robust baseline for future research in the field of MTMCT. A key innovation within this framework is the introduction of a matching cascade with a correction mechanism, which effectively addresses the challenge of erroneous associations from previous time steps. Moreover, the work has successfully integrated three efficient approaches for determining plausible world coordinates from bounding box annotations, demonstrating the practical implications of these methodologies for enhancing real-time tracking accuracy. The competitive results achieved in Track 1 of the 2024 AI City Challenge [27] underscore the framework's potential and validate its performance against other online and offline MTMCT algorithms.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 2

[2] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 7

[3] Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, and Shang-Hong Lai. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10051–10060, 2023. 2

[4] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 2

[5] Nils Friederich and Andreas Specker. Security fence inspection at airports using object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 310–319, 2024. 3

[6] Bipin Gaikwad and Abhijit Karmakar. Smart surveillance system for real-time multi-person multi-camera tracking at the edge. *Journal of Real-Time Image Processing*, 02 2021. 2

[7] Y. He, J. Han, W. Yu, X. Hong, X. Wei, and Y. Gong. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 576–577, 2020. 2

[8] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, and J.-N. Hwang. Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 416–424, 2019. 2

[9] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023. 3, 6

[10] Hyeonchul Jung, Seokjun Kang, Takgen Kim, and HyeongKi Kim. Conftrack: Kalman filter-based multi-person tracking by utilizing confidence score of detection box. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6583–6592, 2024. 2, 3, 7

[11] P. Köhl, A. Specker, A. Schumann, and J. Beyerer. The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1042–1043, 2020. 2

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[13] Jincheng Lu, Meng Xia, Xu Gao, Xipeng Yang, Tianran Tao, Hao Meng, Wei Zhang, Xiao Tan, Yifeng Shi, Guanbin Li, et al. Robust and online vehicle counting at crowded intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4002–4008, 2021. 2

[14] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 6

[15] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023. 1, 2

[16] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Alice Li, Shangru Li, and Rama Chellappa. The 6th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3347–3356, June 2022. 2

[17] Duy MH Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, and Paul Swoboda. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2022. 2

[18] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 588–589, 2020. 2

[19] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6036–6046, 2018. 2

[20] Andreas Specker and Jürgen Beyerer. Toward accurate online multi-target multi-camera tracking in real-time. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 533–537, 2022. 2

[21] Andreas Specker and Jürgen Beyerer. Reidtrack: Reid-only multi-target multi-camera tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5441–5451, 2023. 2, 3

[22] Andreas Specker, Lucas Florin, Mickael Cormier, and Jürgen Beyerer. Improving multi-target multi-camera tracking by track refinement and completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3198–3208. IEEE, 6/19/2022 - 6/20/2022. 2

[23] Andreas Specker, Lennart Moritz, Mickael Cormier, and Jürgen Beyerer. Fast and lightweight online person search for large-scale surveillance systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 570–580, 2022. 3

[24] Andreas Specker, Daniel Stadler, Lucas Florin, and Jurgen Beyerer. An occlusion-aware multi-target multi-camera tracking system. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 4173–4182, 2021. 2

[25] Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Earlybird: Early-fusion for multi-view tracking in the bird's eye view. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 102–111, 2024. 2

[26] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv:1706.06196*, 2017. 2

[27] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024. 2, 5, 8

[28] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE Int. Conf. Image Process.*, pages 3645–3649, 2017. 2

[29] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4256–4265, 2016. 2

[30] Jin Ye, Xipeng Yang, Shuai Kang, Yue He, Weiming Zhang, Leping Huang, Minyue Jiang, Wei Zhang, Yifeng Shi, Meng Xia, et al. A robust mtmc tracking system for ai-city challenge 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4044–4053, 2021. 2

[31] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 7

[32] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 6

[33] Xindi Zhang and Ebroul Izquierdo. Real-time multi-target multi-camera tracking with spatial-temporal information. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. 2

[34] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2, 7