

GRAP-MOT: Unsupervised Graph-based Position Weighted Person Multi-camera Multi-object Tracking in a Highly Congested Space

Marek Socha, Michał Marczyk, Aleksander Kempski, Michał Cogiel, Paweł Foszner, Radosław Zawiski, Michał Staniszewski

Abstract—GRAP-MOT is a new approach for solving the person MOT problem dedicated to videos of closed areas with overlapping multi-camera views, where person occlusion frequently occurs. Our novel graph-weighted solution updates a person’s identification label online based on tracks and the person’s characteristic features. To find the best solution, we deeply investigated all elements of the MOT process, including feature extraction, tracking, and community search. Furthermore, GRAP-MOT is equipped with a person’s position estimation module, which gives additional key information to the MOT method, ensuring better results than methods without position data. We tested GRAP-MOT on recordings acquired in a closed-area model and on publicly available real datasets that fulfil the requirement of a highly congested space, showing the superiority of our proposition. Finally, we analyzed existing metrics used to compare MOT algorithms and concluded that IDF1 is more adequate than MOTA in such comparisons. We made our <https://gitlab.qsystems.pro/publicly/grap-mot> along with the acquired <https://doi.org/10.5281/zenodo.14526116> publicly available.

Index Terms—Detection, multi-camera multi-object tracking, and recognition of objects, image, and video analysis.

I. INTRODUCTION

Monitoring human activity across multiple overlapping camera views in closed environments presents unique challenges and opportunities. While a single camera may suffer

This work was partly supported by European Union Funds Awarded to Bles Sp. z o. o. “Development of a system for analyzing vision data captured by public transport vehicles interior monitoring, aimed at detecting undesirable situations/behaviors and people counting (including their classification by age group) and the objects they carry” under Grant POIR.01.01.01-00-0952/20-00; and in part by the Silesian University of Technology grant for Maintaining and Developing Research Potential under Grant 02/070/BK24/0052 (MM) and 02/090/BK24/0043 (MS). Corresponding author: Michał Staniszewski.

Marek Socha is with the Department of Data Science and Engineering, Silesian University of Technology, 44-100 Gliwice, Poland.

Michał Marczyk is with the Department of Data Science and Engineering, Silesian University of Technology, 44-100 Gliwice, Poland, and also with the Yale Cancer Center, Yale School of Medicine, New Haven, CT 06510 USA.

Aleksander Kempski, Paweł Foszner, and Michał Staniszewski are with the Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, 44-100 Gliwice, Poland (e-mail: mstaniszewski@polsl.pl).

Aleksander Kempski is with the Department of Systems Biology and Engineering, Silesian University of Technology, 44-100 Gliwice, Poland.

Radosław Zawiski is with the Department of Automatic Control and Robotics, Silesian University of Technology, 44-100 Gliwice, Poland.

Michał Cogiel is with Bles Sp. z o. o., 44-100 Gliwice, Poland and QSystems.pro Sp. z o. o., 41-907 Bytom, Poland

Manuscript received XX; revised XX

from occlusions or limited visibility, overlapping fields of view allow for improved robustness and continuity in tracking. Multi-camera multi-object tracking (MOT) enables the integration of fragmented observations into coherent trajectories, which is essential for consistent scene understanding and downstream analytics in applications such as security monitoring, flow analysis, or resource optimization. Our work addresses this problem by proposing a method that operates effectively in such constrained settings without relying on prior identity information or full-body visibility.

Thanks to the increased amount of captured video recordings, the algorithms for person detection, recognition, or tracking are currently developing rapidly. Person detection refers to denoting an individual with a rectangle [1], while tracking refers to assigning a unique identifier (ID) to a rectangle while ensuring that between subsequent video frames the identifier remains constant [2]. This approach can be extended to the multi-camera multi-object tracking, where more than one camera is incorporated in the tracking process. As the image sources are different, to match objects properly across cameras, more effort has to be put in. Due to occlusions, non-overlapping fields of view (FOV), changes in angle, and lighting, this problem remains a challenge. A related task is person re-identification, which focuses on retrieving instances of the same person across different cameras, typically under non-overlapping FOV conditions. However, unlike re-identification, MOT does not assume the prior availability of gallery images and often deals with overlapping camera views. Additionally, MOT requires not only maintaining consistent identity labels across time within a single camera view but also ensuring correct identity matching across different cameras.

In the given work, we introduce a novel method named GRAP-MOT for tracking many persons in a closed area using at least 2 different cameras with overlapping views, placed to cover all areas of the space from different angles. The multi-camera multi-object tracking (MOT) task in such an environment is problematic due to numerous occlusions, changes in viewing angle, and strongly differing person detection sizes. Using an internal dataset where the scene is usually highly condensed, has limited time for data acquisition, and the persons tend to stand still during records, we provided a deep analysis of each element of the MOT problem to choose the final solution. Thus, we proposed a complete system capable of tracking individual persons on a single camera and tracking multiple persons when many cameras are used. We tested our

solution on a specially prepared mock-up simulating a closed space with a different number of people, as well as on available datasets meeting the assumed requirements. Our source code¹ along with the acquired data² are publicly available. The main contributions of this paper are summarized as follows:

- In contrast to other MOT methods, GRAP-MOT relies on the strengthening of the tracklets (i.e. a sequence of short detections) association over time rather than assuming immediate proper track matching. It makes convergence longer but gives higher robustness of detections.
- GRAP-MOT relies only on constant observation of the weighted graph and ongoing updating of ID numbers using acquired tracks, information about a person's position, and their characteristic features. No other data are needed.
- Based on the information from the graph, the method uniquely analyzes possible connections and creates object-tracking groups based on their similarities.
- A module for estimating people's positions, not available elsewhere, is a key element of multi-camera multi-object tracking.
- GRAP-MOT requires little to no supervision. The only trained model is the one for feature extraction, which could also be acquired from public repositories.
- Additionally, we conducted an important analysis of comparative metrics for assessing the effectiveness of MOT, proposing the IDF1 against the MOTA, which frequently appears in other works.

A. Review of existing tracking methods

Person detection and tracking methods are well-developed in the literature (Table I). Early works in the area of person detection consisted of either simple shape delineation methods based on active contours [3], [4] or face detection by image features [5]. Since 2015, there has been a rapid rise in interest regarding detection methods driven by the development of new deep learning models [6]. The most popular detection networks are R-CNN [7], Faster R-CNN [8], Mask R-CNN [9], and YOLO [10] method with its multiple versions. Tracking aids detection by allowing one to follow objects across video frames, thus bridging the gaps between detections and frames. In 2006 the tracking methods were split into multiple categories [11] including point-based tracking [12], [13] where a point in space represents the tracked object, kernel-based methods [14], [15] referring to object shape and motion and silhouette tracking [16], [17] where the object contour and image features are the matching factor. Currently, the most common subjects for detection and tracking are rectangular detections, so-called bounding boxes, for which point-based tracking methods are particularly popular. Usually, those methods are based on the Kalman Filter [12] due to its simplicity but high efficiency. Methods like POSE [18] or SORT [19] utilize the filter to predict the next position solely based on the current position and motion. Later, hybrid systems emerged, connecting the point-based and silhouette categories. Methods like DeepSORT [20], [21], ByteTRACK

[22], or MOTDT [23], which are currently regarded as the state-of-the-art, use deep learning methods to define object features and employ the information from the Kalman filter to support their decision.

Multi-camera multi-object tracking (MOT) ensures that across multiple cameras, detection of the same person or object shares the same identifier [24]. These methods can be divided according to the location, the analyzed object, the distribution of the objects, the number of cameras, the degree of camera overlap, and the timing of the recording. In recent years, we have witnessed an increase in the popularity of the MOT methods development, mostly due to the AI city challenge [25], [26], resulting in many works centered around the tracking of motor vehicles. The deep learning-based solutions are predominant in this area. For example, researchers used them to extract features of the objects and employed GPS location to re-identify and track cars [27], [28]. Common practice is to employ the cars' trajectories to better match tracks between cameras [29]. Found tracks can be matched to the reference camera [30] or analyzed simultaneously using, for example, graph methods [29], [31], [32].

The problem of tracking people in a highly congested closed space, where occlusion frequently appears and there is higher variation in detection sizes, is much harder than tracking in an open area. There is a limited number of works covering MOT in these types of scenes. For example, MOT in the operating room uses skeletal pose estimation with tracking to compensate for the lack of colorful clothes and visible faces [33]. The method for person tracking in the warehouse employs elaborate image processing to deal with the sudden changes in the lighting due to the large windows [34], [35]. To deal with the person's tracking in the shop, the authors used the approach revolving around the density maps since the camera was placed on the ceiling [36]. The same authors proposed later a similar approach, but with the addition of a trajectory showing the improvement over previous work [37]. Most recent approaches tend to use the spatial and temporal graph approach, where the spatial graph matches tracks on the n th frame across different views and temporal associates objects between frames. For example, the DyGLIP method [38] uses a dynamic graph network with attention to match spatial information and then match it with temporal tracks. The ReST separates these tasks into two, separately trained graphs; the first graph combines spatial information, and the second graph has the task of combining tracks in time between successive frames [39].

The problem of person re-identification, which focuses on comparing each probe image to a predefined set of gallery images and selecting the most similar match, is also well described. In [40], person re-identification (Person ReID) is understood as the task of matching individual people from images collected from multiple non-overlapping cameras. The process in practice involves comparing a query image with an image gallery, and this paper is a review of methods based on deep learning models to deal with domain shifts, including clothing changes, in Lifelong ReID scenarios. The re-identification task is defined in the same way in the case

¹<https://gitlab.qsystems.pro/publicly/grap-mot>

²<https://doi.org/10.5281/zenodo.14526116>

of the work [41], but is considered in the context of multiple heterogeneous (airborne and ground-based) cameras. Through this, a view-difference arises, and in response, the authors propose a View-decoupled Transformer (VDT) model, based on the Transformer architecture (more precisely ViT-Base), which aims to separate view-related features from features crucial for identity identification. In this paper, the authors used two datasets containing both ground and aerial views, their proprietary synthetic CARGO (Civic AeRial-GrOund) dataset and the actual AG-ReID dataset. The work [42] focuses on the problem of text-image person re-identification (TIReID), the understanding of which is different from previous work. Re-identification is defined as the task of finding an image of a person in a large image gallery based on a textual description. The main challenge addressed in the paper is the problem of noisy correspondence (NC) in training data, and the authors propose a Robust Dual Embedding (RDE) method that uses OpenAI's CLIP-B/16 pre-trained model as the underlying encoders for learning robust visuo-semantic associations. The work [43] considers the problem of re-identifying persons (Person ReID), relative to image galleries based on different types of input data such as image, text, or sketch. To make this possible and to increase generalisation, the authors propose an AIO model that uses a frozen, pre-trained base model based on the Vision Transformer (ViT) architecture as a shared feature encoder. The work uses both real and synthetic collections. Due to the lack of a large amount of annotated data, sketches are sometimes generated from RGB images. The work performs evaluations on independent collections such as Market1501 and CUHK-PEDES, for example. The paper [44] also considers re-identification as a probe image search in a gallery, but considers it in the context of Lifelong Person Re-identification - LReID. It proposes a Distribution-aware Knowledge Prototyping (DKP) method, the premise of which is not to store learning data, bypassing the privacy issues and computational costs associated with memory-based methods. Instead, it creates prototypes to store this information, creating a distribution for each sample, taking into account its individual variability. The method is trained on Market1501, DukeMTMC-reID, CUHK-SYSU, MSMT17-V2 and CUHK03 image data, and is tested on non-dependent sets. In the work [45], re-identification is also referred to as an image gallery search process. The main innovation of the work is to propose a new loss function, called Differentiable Retrieval-Sort Loss (DRSL), to optimise the feature distribution, which is used with typical re-ID model architectures such as ResNet-50 used as the core of the network. The work performs evaluations on sets such as Market1501, CUHK03, and MSMT17, for example. The paper [46] introduces the concept of re-identification as Text-to-Image Vehicle Re-Identification, which is understood as searching for a target vehicle by matching a text description with images in a photo gallery. The context is urban surveillance systems using multiple non-overlapping cameras. To reduce the gap between modalities (text and image), the authors propose a Multi-scale multi-view Cross-modal Alignment Network (MCANet), which uses ResNet-50 for vision and BERT with text convolution modules for text as the core. The work [47] focuses on the re-

identification of people visible on non-overlapping cameras based on image galleries. However, in addition to using RGB images for this purpose, the authors additionally use depth images of images. To reduce the significant discrepancies between RGB modality and depth, the authors propose an Intermediary-Generated Bridge Network (IBN), a network using a ResNet50-based architecture as the core, complemented by a multi-modal transformer and circle contrast learning module. In this work, the authors use real data from the RobotPKU, BIWI, and SYSU-MM01 collections.

The idea of using graphs for classic re-identification was used previously in the literature. However, since all of these methods assume that there is a pre-defined reference database of known objects, they cannot be used in the MOT task defined above. In [48], person re-identification is formulated as the task of locating a target individual in an image gallery by comparing a probe image against all gallery images. While most existing methods rely solely on probe-to-gallery (P2G) similarities, the proposed Deep Group-shuffling Random Walk Network incorporates both P2G and gallery-to-gallery (G2G) similarities into an end-to-end learning framework. The model is based on a Siamese CNN architecture using ResNet-50 as the backbone, and its effectiveness is demonstrated on standard multi-camera datasets: Market-1501, CUHK03, and DukeMTMC. Similarly, [49] defines re-identification as matching a probe image within a gallery. The proposed Similarity-Guided Graph Neural Network (SGGNN) integrates both P2G and G2G relationships to refine similarity estimation during both training and inference. The method also builds on a Siamese CNN framework and is evaluated on Market-1501, CUHK03, and DukeMTMC. In [50], the re-identification task focuses on vehicles across non-overlapping camera networks. The approach uses ResNet-50 for initial feature extraction and introduces a Camera Topology Graph Convolutional Network (CT-GCN), which explicitly models spatial and directional relationships between cameras. The graph is constructed with four hierarchical levels: system-wide (global connection), position (spatial proximity), orientation (directional alignment), and individual (self-loop). This structure allows the model to learn more discriminative, camera-independent features. The method in [51] addresses person re-identification across disjoint camera views using a Masked Graph Attention Network (MGAT). It employs ResNet-50 for feature extraction and leverages a graph attention mechanism to model global relationships among all gallery images. This enables more effective refinement of features for probe-gallery matching. The approach is trained and tested on iLIDS-VID, PRID2011, MARS, and Market-1501 datasets. Finally, [52] tackles vehicle re-identification by defining it as an unguided search task across multiple camera views. The method uses OpenAI's CLIP-B/16 model to extract joint visual and textual features and constructs a graph (VLCGT) to model relationships among training samples. This graph is incorporated into the learning process to improve similarity estimation and enhance feature discriminability.

TABLE I
SUMMARIZATION OF PAPERS SIMILAR TO OUR PROBLEM. P:
PEOPLE/PERSON MOT, CS: CLOSED SPACE, HD: HIGH DENSITY, OV:
OVERLAPPING VIEWS, CA: CODE IS AVAILABLE

Ref	Year	P	CS	HD	OV	CA
[39]	2023	✓	-	✓	✓	✓
[53]	2023	✓	-	-	-	-
[35]	2023	✓	✓	-	✓	✓
[54]	2023	✓	-	-	-	-
[55]	2023	✓	✓	✓	✓	-
[33]	2022	✓	✓	-	✓	-
[28]	2022	✓	-	✓	-	-
[27]	2022	-	-	-	-	✓
[31]	2022	-	-	-	-	✓
[56]	2021	✓	-	-	-	✓
[38]	2021	✓	-	-	✓	✓
[37]	2020	✓	-	-	✓	✓
[36]	2020	✓	✓	✓	-	-
[34]	2020	✓	✓	-	✓	-
[29]	2020	-	-	-	✓	-
[30]	2020	-	-	-	✓	-
[57]	2020	✓	-	-	-	✓
[58]	2019	✓	-	-	-	✓
[59]	2018	✓	-	-	✓	✓
[60]	2018	✓	-	-	-	-

II. MATERIALS AND METHODS

A. Internal Dataset

A closed-area model was constructed in the laboratory to acquire an internal benchmark dataset. Three cameras were installed, including two standard cameras in adjacent corners of the model (Figure 1a, Figure 1c) and one fish-eye camera (Figure 1b) in the center near the entrance. Such camera placement allows the entire area inside the model to be visible without leaving blind spots. A total of 14 scenes were recorded, capturing the behaviour of numerous people, ranging from 2 to 15 people per recording. Each scene was created in a closed-area model and comprises 10 video recordings lasting about 10 seconds, and 252 frames on average. Using a detection model based on the YOLOv7 [61] architecture, persons' heads were found and surrounded by bounding boxes to support bounding box labeling for tracking. Head detections were curated by manually correcting missing detections, adjusting the bounding box area to the persons' heads, and removing incorrect detections. After data curation, detections were labeled to match across all cameras.

B. External datasets

The CAMPUS [62] public dataset was used as an external validation set. The head detections and annotations made in the study of Yuan Xu *et al.* [62], [63] were used. CAMPUS dataset [62] contains 4 recordings with a partially overlapping field-of-view (FOV) taken from 4 cameras, each with a resolution of 1980x1020, and recorded at a 30 fps frame rate. The Auditorium subset comprises footage captured by two cameras placed at the entrance and two within the auditorium, amounting to about 5,000 frames. The Garden1 and Garden2 subsets feature recordings from four cameras in the park, each with overlapping fields of view, containing 3,000 and 6,000 frames, respectively. The Parkinglot subset includes videos

from four cameras situated in various spots in the parking area, also with overlapping fields of view, comprising 6,500 frames.

C. Head detection method

YOLO (You Only Look Once) version 7 [61] was trained to detect human heads on images from a closed-area model. According to our previous work [64], head detections are more robust to occlusions and more stable in comparison to full-body detections. From the available models, the YOLOv7-X was chosen as a compromise between precision and evaluation time for the head detection task. In training, default parameters were used together with a batch size equal to 8 and the number of epochs equal to 300. Two datasets were used to train the model: (i) the CrowdHuman benchmark dataset [65]; (ii) our internal dataset of images taken in the closed-area model [64]. The CrowdHuman consists of 470 thousand human unique instances with an average of 23 persons per image and different levels of occlusions. From this database 15,000, 4,370, and 5,000 images were used for training, validation, and testing, respectively. From our internal dataset, 543 images were used for model training and 61 for model validation.

D. Tracking methods

To track people within a single camera the following methods were evaluated: SORT [20], DeepSORT [21], FairMOT [66], and ByteTrack [22]. These methods follow similar steps of analysis. Given the bounding box from the object detection method, the first step is to predict the next position of the tracking object. Secondly, there is a data association phase where predicted positions are matched with similar previous positions, creating a tracklet, i.e. a sequence of short detections (in contrast to track, which is a whole trajectory of an object). Finally, there is a track management phase where the tracks are updated, added, or deleted.

Every tested tracking method follows the presented workflow but each provides some novel tracking solution. The oldest method SORT is a point-tracking algorithm based on a Kalman filter which predicts the object's future location using only the current state (position and velocity). The association phase uses the Hungarian algorithm to match tracklets with predicted positions based on the bounding boxes' Intersection over Union (IoU). The DeepSORT method was created to enhance the association step by adding the appearance features extracted by the deep learning model tailored by the user to match the tracked object. Matching is done by combining IoU and cosine similarities between the appearance features. FairMOT attempts to unify position detection and feature extraction into one network, making it efficient but less flexible. ByteTrack introduced a novel way of dealing with low-confidence detections where such detections are matched during the association phase with the existing tracklets making them more likely to appear in the future. Additionally, low-confidence detections are used to prevent loss of the tracklet during occlusions.

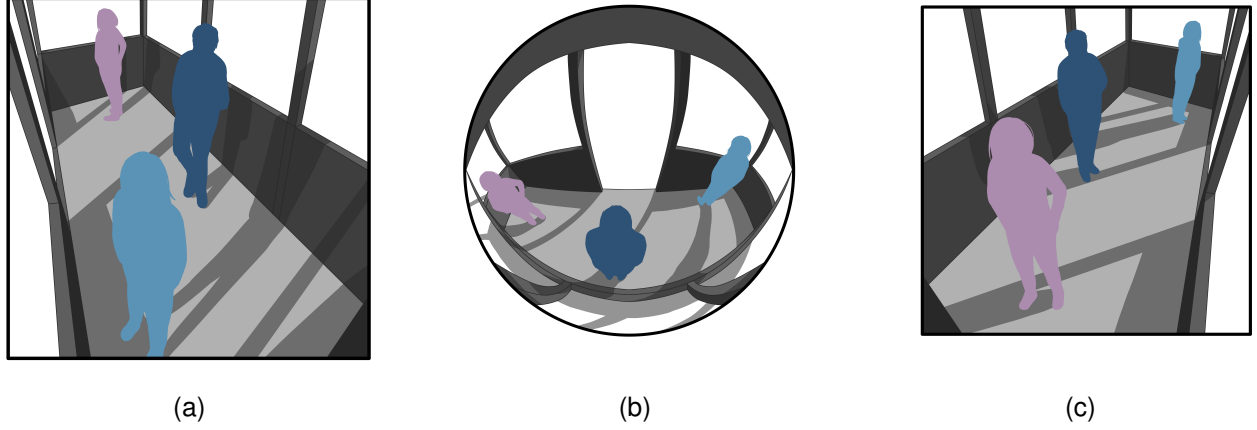


Fig. 1. Iconographic visualization of camera distribution on the closed-area model frame. The exemplary person's multi-object tracking on different camera views is included in different colors, where left/right (a and c) are typical views and (b) in the middle is a fish-eye type view.

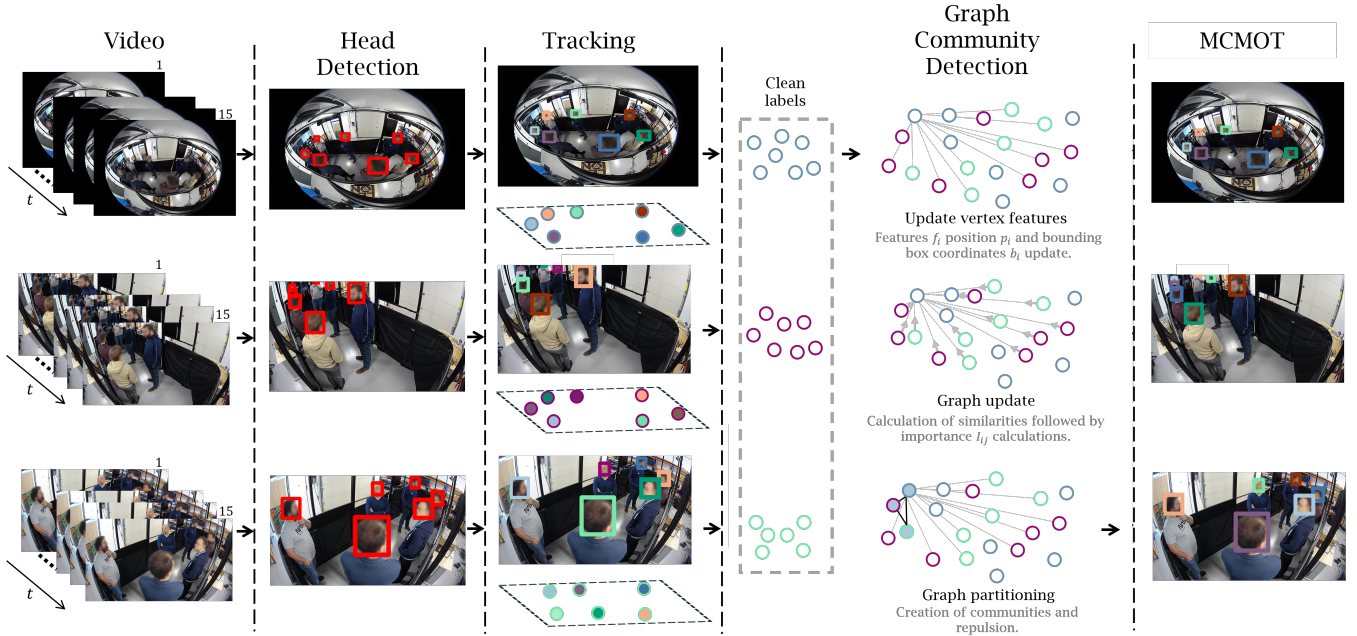


Fig. 2. Overview of the GRAP-MOT Multi-Camera Multi-Object Tracking Pipeline. Video frames from multiple cameras are processed sequentially. A head detection module outputs bounding box coordinates for each detected individual. Within each camera view, a tracking algorithm assigns temporary identity labels to detections, forming short-term trajectories (tracklets). A position estimation module projects these tracklets onto a common spatial grid. All tracklets are then represented as nodes in a graph, where edges connect tracklets originating from different cameras. An importance coefficient is computed for each edge to quantify the likelihood of cross-camera identity correspondence. Tracklets are clustered across cameras into groups, with group sizes limited by the total number of cameras. Finally, unified identity labels are assigned within each group, yielding consistent tracking across all camera views.

E. Feature Extraction

Different convolutional neural networks were tested for feature extraction from head images, namely OSNet [67], ONet-AIN [68], and ResNet50 [69], which were trained for the MOT task on the Market1501 [70] and DukeMTMC [71] collections with the torchreid repository [72]. The training set included all images from the collections mentioned above (see description of datasets). The networks were trained with 100 epochs at max, with input sizes set to 256x180 and images augmented using random flip and random crop methods. Head images were usually smaller than the desired size so they were up-scaled using nearest-neighbours interpolation method.

F. Position estimation

The XGBOOST [73] model was used to estimate a person's position, which was originally used in [64] to predict the X and Y coordinates of the person's position in a bus, and also for the decision if each person is inside or outside of the bus. The most important parameters were tuned using the Bayesian optimization method [74]: (i) eta - step size shrinkage used in the model update (range 0.001-0.5); (ii) max depth - maximum depth of a tree (range 1-20); (iii) gamma - minimum loss reduction required to make a further partition on a leaf vertice of the tree (range 0-0.1); (iv) colsample bytree - subsample ratio of columns when constructing each tree (range 0.4-1); (v)

min child weight - minimum sum of instance weight needed in a child (range 0.1-10); (vi) subsample - subsample ratio of the training instances that occur once in every boosting iteration (range 0.5-1); (vii) lambda - L2 regularization term on weights (range 0-10); (viii) alpha - L1 regularization term on weights (range 0-10). Selection of the best parameters was done using 10-fold cross-validation. The position estimation was calculated only for the internal dataset. In the case of the external datasets, the tracklets did not have this attribute.

G. Graph update

The cameras are denoted as c , and tracklets are denoted as T . The number of cameras is constant and equal to C . The number of tracklets per camera may not be equal. Let us denote the total number of tracklets as M and treat them as a single set, despite their different camera origins, for simplicity. The method assumes that the tracklets contain extracted object features (f), object position estimation (p), detection bounding box coordinates (b), and the camera of the origin (c) (Equation 1).

$$T_i = \{f_i, p_i, b_i, c\} \quad c \in [1..C] \quad i \in [1..M] \quad (1)$$

To combine tracklets from multiple cameras, a graph defined by the set of vertices V (Equation 2) and the set of edges E (Equation 3) is constructed. In vertices, tracklet information is stored. Only vertices with tracklets that originate from different cameras are connected by edges (tracklets from the same camera are not connected). Thus, edges represent the inter-camera relation between the tracklets.

$$V = \{v_1, v_2, \dots, v_M\} \quad (2)$$

$$E = \{\{v_i, v_j\} \mid i \neq j \wedge c_i \neq c_j\} \quad i, j \in [1..M] \quad (3)$$

After creating the graph, the cosine distance between the object features and between the position estimates is calculated for each edge. The resulting distances are normalized to the range from 0 to 1 using the sigmoid (σ) function and converted to similarities for features s_{ij}^f and for positions s_{ij}^p . Additionally, in each edge, occurrences (o_{ij}) are initialized with the value 1. The occurrence parameter counts the number of video frames in which two vertices connected by the edge were labelled as the same person.

$$s_{ij}^f = 1 - \sigma \left(\frac{f_i \cdot f_j}{\|f_i\|_2 \|f_j\|_2} \right) \quad i \neq j \wedge c_i \neq c_j \quad i, j \in [1..M] \quad (4)$$

$$s_{ij}^p = 1 - \sigma \left(\frac{p_i \cdot p_j}{\|p_i\|_2 \|p_j\|_2} \right) \quad i \neq j \wedge c_i \neq c_j \quad i, j \in [1..M] \quad (5)$$

To quantify the inter-camera relation between tracklets, for each edge, the importance value I is calculated. The edge importance value is updated at each frame and is defined by the equation (Equation 6). The larger the I on the edge between the vertices, the more probable they correspond to the same

person. Division by three keeps the I update value in the range of 0 and 1, where auxiliary frame-related index t is used to indicate consecutive updates of the I parameter.

$$I_{ij} = I_{ij}^{t-1} + \frac{1}{3} \left(s_{ij}^p + s_{ij}^f + \frac{o_{ij}}{\sum_{l=1}^{M-1} \sum_{k=l+1}^M o_{lk}} \right) \quad (6)$$

If the position estimates are unavailable, a different form of the equation to calculate edge importance (Equation 12) is used.

Let $K = \{k_1, k_2, \dots, k_{M_k}\}$ and $L = \{l_1, l_2, \dots, l_{M_l}\}$ represent vertices (tracklets) of the graph from the two different cameras. For each vertex k_i (tracklet), the cosine distance between the bounding box coordinates of k_i and other vertices from the same camera is calculated. This creates an intra-camera neighbour vector INV (Equation 8). The operation is repeated for each camera. The goal is to compare the values of the vector INV between vertices from different cameras. Thus, each vector INV of K camera vertices is compared to each vector INV of L camera vertices (Equation 9), giving IRM matrix. Finally, the bounding box relation coefficient r , which reflects neighbourhood similarities of the tracklets between cameras, is calculated as the arithmetic mean of the previous value (i.e. initial value or previous frame value) and the sum of IRM .

$$INV_k = \frac{b_{k_i} \cdot b_{k_j}}{\|b_{k_i}\|_2 \|b_{k_j}\|_2} \quad i \in [1..M_k] \quad (7)$$

$$INV_k = [INV_{k_1}, INV_{k_2}, \dots, INV_{M_k}] \quad (8)$$

$$IRM_{k_i l_j} = (1 - |INV_{k_i} - INV_{l_j}'|)^2 \quad i \in [1..M_k] \quad j \in [1..M_l] \quad (9)$$

$$r_{ij}^t = \frac{r_{ij}^{t-1} + \sum IRM_{ij}}{2} \quad (10)$$

The calculated bounding box relation coefficient r is then used to calculate the edge importance using the alternative equation that relies more on the similarity between features s^f and has r as the support for the decision (Equation 12). For this purpose, the smoothed occurrences q_{ij} coefficient is defined (Equation 11) where α is the smoothing factor set to 50 for the analyzed problem. Division by $1 + o_{ij}$ scales the I update value to the range 0 and 1, assuring numerical boundness.

$$q_{ij} = 1 - \exp\left(\frac{-1}{\alpha} o_{ij}\right) \quad (11)$$

In the proposed importance equation, bounding box features are the main anchor. The bounding box relation coefficient comes online after the anchor was established, meaning the occurrence value was updated for a given connection. Only then is q different from 0.

$$I_{ij}^t = I_{ij}^{t-1} + \frac{r_{ij} q_{ij} + s_{ij}^f}{1 + o_{ij}} \quad (12)$$

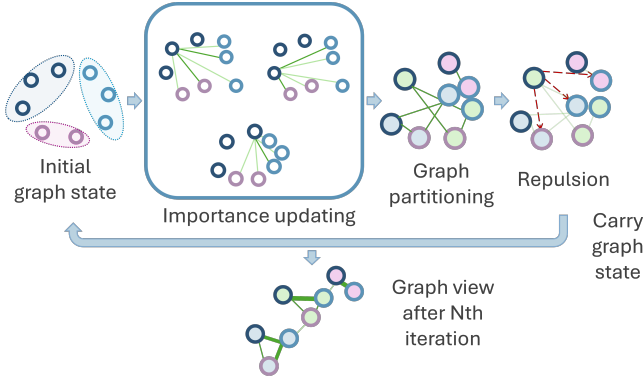


Fig. 3. Overview of graph update, partitioning and repulsion processes.

H. Graph partitioning

To organize the vertices in groups, communities were created. Communities are groups of points, in this case, vertices, which commonly interact with each other. The interaction between the vertices is simulated in the presented method by the Importance value stored in the edges of the graph. For this purpose, the greedy modularity communities detection algorithm was used [75]. The algorithm was modified to limit the maximum number of vertices in the society to the number of cameras used in the MOT task. The modularity measure is used to quantify the graph community partitioning quality. Given the list of communities, it rates them based on the number of connections inside the community. The equation also has a weighted variation (Equation 13) where a given edge feature present between the vertices in the community replaces several connections. In both cases, classic and weighted versions, modularity is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \quad (13)$$

where m is the sum of the edge weights, A is the adjacency matrix, γ is the resolution parameter, k is the weighted degree of i (or j), and $\delta(c_i, c_j)$ is a Kronecker delta (if i and j are in same community 1, if are in different 0). The γ resolution parameter determines whether intra-group or inter-group relationships are more important.

After the communities were created, they were validated by checking whether each community had tracklets from the same cameras. In such an event, vertices in the validated community are repulsed from each other by subtracting the Repulsion coefficient e (Equation 14) of the current frame from the stored Importance value, making it less probable for vertices to join each other communities in the future. The repulsion coefficient is defined for the vertex as the sum of the neighboring edge occurrences divided by the number of neighbors (N_n).

$$e = \frac{\sum_{i=1}^{N_n} o_i}{N_n} \quad (14)$$

$$I_{ij} = I_{ij}^{t-1} - e \quad (15)$$

I. Evaluation metrics

There are two groups of commonly used metrics for multi-camera multiple-object tasks, the CLEAR-MOT group [76], [77] and the ID group [78]. The former focuses mainly on assessing tracking and detection quality, and the latter on assessing match quality. From these groups, the two most important metrics for the MOT task were selected. MOTA (Multiple Object Tracker Accuracy) measures overall tracking and detection accuracy, considering missed detections, mismatches, identity switches, and false positives. A high MOTA value indicates good tracking quality. IDF1 measures the quality of identifier assignment for tracks. It takes into consideration both precision and recall of identities across frames. High IDF1 indicates that appropriate identifiers are maintained throughout the frame sequence.

The pymotmetrics package [79] was used to calculate the listed metrics. To calculate them for multi-camera multiple-object tasks, each camera was assigned an identifier from 0 to N_c (number of cameras). Detection results were added sequentially frame by frame. Since the frame number has to be unique, even if the results come from different cameras, the frame number was multiplied by 1000, and the camera ID was added. In this way, MOTA and IDF1 could be calculated for a multi-camera problem.

J. Statistical analysis

Multiple tracking methods, feature extraction, and community detection were tested during experiments. To compare the difference between the results of different methods, the Kruskal-Wallis test was used. In all tests, the statistical significance level was set to $\alpha = 0.05$. If there was enough evidence to state the inequality of medians, the test was followed by Nemenyi post-hoc test. If during the test of method 1 against method 2, the p-value was smaller than the significance level and the median result of method 1 was greater than the median of method 2, method 1 was considered better.

III. RESULTS

In our experiments, the baseline configuration combined DeepSORT for tracking, ResNet50 for feature extraction, and greedy modularity optimization for community detection. We then evaluated the impact of changing each module individually and of disabling the position estimation module. Median IDF1 and MOTA metrics were computed per dataset to compare configurations. The optimal combination was subsequently benchmarked against other methods on the external dataset.

A. Performance of the tracking methods in a congested space

At first, we analyzed the performance of different object tracking methods on a single camera for the MOT problem in the highly congested space. IDF1 and MOTA metrics were calculated for 140 recordings, and metrics were aggregated by median value based on the number of people present in the scene (Supplementary Table 1, Supplementary Figure 1). In Table II, we present the median, mean, and standard deviation

TABLE II

TRACKING METHODS' PERFORMANCE COMPARISON ON THE INTERNAL DATASET. THE MEDIAN, MEAN AND STANDARD DEVIATION OF THE CREATE GROUP METRICS ARE SHOWN IN THE TABLE.

Tracker	Metric	Aggregation		
		Median	Mean	STD
DeepSORT	IDF1	79.279	77.608	14.701
	MOTA	63.689	64.202	21.339
SORT	IDF1	81.492	79.975	15.115
	MOTA	72.269	72.950	18.938
FairMOT	IDF1	75.799	74.021	17.189
	MOTA	70.506	70.110	20.228
ByteTrack	IDF1	80.969	78.980	15.074
	MOTA	70.037	71.414	20.285

of the results for each tracking method. We found that the obtained IDF1 metrics are statistically different between methods (p -value=0.0245, $\alpha = 0.05$). To find the best-performing method, we confronted post hoc Nemenyi test results with the median of the sets' scores. According to the analysis, the best tracking method is SORT, since it is better than all other methods (relative to DeepSORT p -value=2.624e-02, FairMOT p -value=4.796e-14, and Byte Track p -value=0.0186). The second best proved to be Byte Track, which was better than DeepSORT and Fair MOT (relative to DeepSORT p -value=0.0262, Fair MOT p -value=7.132e-11). The DeepSORT method, which was a default method for the experiments, placed third, proving to be better than Fair MOT (relative to Fair MOT p -value=3.991e-04).

B. Searching for the best model architectures for a person's head image

After selecting the most effective single-camera tracking method, we tested the performance of different neural network models for feature extraction from a person's head bounding box. IDF1 and MOTA were aggregated by a median concerning their specific set (Supplementary Table 2, Supplementary Figure 2). In Table III, we present the median, mean, and standard deviation of the results for each neural network model used for feature extraction. We found statistical differences between the results of different models. After the post hoc Nemenyi test and median result analysis, the best-performing network proved to be the ResNet50 model (relative to OSNet x1 p -value=4.4409e-14, OSNet x0.5 p -value=6.05071e-14, OSNet AIN x1 p -value=5.773e-14 and OSNet IBN p -value=7.549e-13). The differences in results of other model architectures were not statistically significant (p -values>0.05). Thus, there is no difference in the median of the quantitative variable across models aside from the ResNet50. The other model architectures were placed equally in last place.

C. Detection of graph communities with position estimation module

Our method revolves heavily around the idea of community detection on the graph. The Importance variable is updated every frame for each connection, and based on the strength of the connection's Importance communities are created. We tested the following algorithms for community detection:

TABLE III

NEURAL NETWORK MODELS' PERFORMANCE COMPARISON ON THE INTERNAL DATASET. THE MEDIAN, MEAN AND STANDARD DEVIATION OF THE CREATE GROUP METRICS ARE SHOWN IN THE TABLE.

Model	Metric	Aggregation		
		Median	Mean	STD
ResNet50	IDF1	79.279	77.608	14.701
	MOTA	63.689	64.202	21.339
OSNetx1	IDF1	63.004	64.768	22.696
	MOTA	52.388	55.848	25.253
OSNet x0.5	IDF1	62.245	64.271	22.895
	MOTA	47.439	55.507	25.827
OSNet AIN x1	IDF1	63.135	63.926	22.015
	MOTA	50.678	54.975	25.828
OSNet IBN	IDF1	65.769	65.368	22.269
	MOTA	52.892	55.078	25.140

TABLE IV

COMMUNITY DETECTION ALGORITHMS' PERFORMANCE COMPARISON ON THE INTERNAL DATASET. THE MEDIAN, MEAN AND STANDARD DEVIATION OF THE CREATE GROUP METRICS ARE SHOWN IN THE TABLE.

Community Detection	Metric	Aggregation		
		Median	Mean	STD
GMM	IDF1	79.279	77.608	14.701
	MOTA	63.689	64.202	21.339
LOU	IDF1	39.454	48.511	26.641
	MOTA	36.354	43.214	24.089
ASYN	IDF1	28.66	33.597	13.141
	MOTA	50.749	52.730	9.491
SC	IDF1	17.048	22.269	12.490
	MOTA	27.888	33.402	24.677
GNI	IDF1	29.322	31.442	12.951
	MOTA	80.464	77.932	10.647
GNC	IDF1	30.063	32.101	17.067
	MOTA	47.374	48.935	12.549

greedy modularity maximization (GMM) [75], Louvain (LOU) [80], asynchronous label propagation (ASYN) [81], spectral clustering (SC) [82] and Girvan–Newman weighted by Importance (GNI) and on betweenness centrality (GNC) [83]. In the case of greedy modularity maximization and Louvain methods, we limited the number of possible community members to the number of cameras.

Again, the IDF1 and MOTA were aggregated by median value based on the number of people present in the scene (Supplementary Table 3, Supplementary Figure 3). In Table IV, we present the median, mean, and standard deviation of the results for each community detection method. We found significant differences between the result groups (p -value=0.00237). The post hoc Nemenyi joint with a median of the sets' aggregations revealed that the best-performing method was GMM (relative to LOU p -value=2.125e-08, ASYN p -value=5.596e-14, SC p -value=3.4e-38, GNI p -value=3.4e-38, GNC p -value=3.4e-38). Other methods' median IDF1 scores largely deviate from the median IDF1 of GMM. The Louvain method was promising at first, showing high IDF1 values with a small number of people (2-3 people); however, as the number of people in the space increased, the quality of MOT declined steeply.

D. Multi-camera multi-object tracking without position estimation module

Position estimation is one of the key features of GRAP-MOT, but since it was tailored to an internal dataset, we

TABLE V

MOT RESULTS WITHOUT POSITION ESTIMATION MODULE; PERFORMANCE COMPARISON ON THE INTERNAL DATASET. THE MEDIAN, MEAN AND STANDARD DEVIATION OF THE CREATED GROUP METRICS ARE SHOWN IN THE TABLE.

Tracker	Metric	Aggregation		
		Median	Mean	STD
DeepSORT	IDF1	36.076	40.350	14.044
	MOTA	11.076	15.638	12.571
SORT	IDF1	31.984	34.529	13.435
	MOTA	18.238	19.159	5.585
ByteTrack	IDF1	32.991	36.034	12.470
	MOTA	14.116	16.119	6.771

also evaluated the system without it. When unavailable, a supplementary bounding box relation module estimates positional similarity across cameras. While the feature extraction and community detection modules show clear advantages, the choice of tracking method remains uncertain; hence, additional experiments were conducted with DeepSORT, SORT, and ByteTrack.

We found significant differences in IDF1 (p-value=0.00237) and MOTA (p-value=8.111e-10) metrics between tracking methods (Supplementary Table 4, Supplementary Figure 4, Table V shows the median, mean, and standard deviation of the metrics for recording groups). The post hoc test revealed that this time the DeepSORT was the best-performing method (p-value=4.832e-10 in comparison to SORT; p-value=8.519e-05 in comparison to Byte Track). There was no statistically significant difference between the results of SORT and Byte Track (p-value=0.0693).

E. GRAP-MOT evaluation on the external dataset

We analyzed the CAMPUS dataset to compare our approach with other existing solutions on publicly available real data. Specifically, we used the Auditorium, Garden1, Garden2, and Parkinglot subsets. Overall, the CAMPUS dataset's tracks are not ideal, as some recordings took place outside, and not all camera views were overlapping. Also, the CAMPUS dataset does not include any information about a person's position; thus, the GRAP-MOT was executed without the position estimation module. We used the best-performing algorithms from the previous experiments (the ResNet50 model and Greedy Modularity Maximisation community detection method) for feature extraction and community detection. The tracking algorithm was chosen based on experiments without the position estimation module (DeepSORT tracking method).

It is impossible to compare IDF1 with other methods because the authors did not provide them in their articles. In terms of MOTA, the GRAP-MOT method is comparable to TRACTA, STP, HCT, KSP, and POM aside from the Garden1 results (Table VI). In comparison with the DyGLIP, the MOTA metric differences are high. However, this result only means that there were not many ID switches, since MOTA does not supply information about the wellness of the label assignment.

Additionally, we compared our method with the ReST model [39]. To get IDF1 and MOTA, we used the existing implementation of the ReST method shared on GitHub. Each recording from the CAMPUS dataset was evaluated with its

TABLE VI

MOT RESULTS WITHOUT POSITION ESTIMATION MODULE; PERFORMANCE COMPARISON ON THE CAMPUS EXTERNAL DATASET. RESULTS ARE GIVEN MAINLY FOR THE MOTA PARAMETER, WHICH IS ONLY AVAILABLE FOR OTHER METHODS.

Method	Metric	Recording			
		Auditorium	Garden1	Garden2	Parkinglot
GRAP-MOT	IDF1	26.69	14.94	19.02	24.55
	MOTA	22.98	18.68	35.29	33.15
DyGLIP	IDF1	-	-	-	-
	MOTA	96.7	71.2	87	72.8
TRACTA	IDF1	-	-	-	-
	MOTA	33.7	58.5	35.5	39.4
STP	IDF1	-	-	-	-
	MOTA	24	57	30	28
HCT	IDF1	-	-	-	-
	MOTA	20.6	49	25.8	24.1
KSP	IDF1	-	-	-	-
	MOTA	17.6	28.1	21.9	14
POM	IDF1	-	-	-	-
	MOTA	16.2	22.4	14	11

respective spatial and temporal graphs. Tests were conducted based on the last 20% of recordings' frames, as the first 80% was used for graphs training. Our method, GRAP-MOT, was evaluated using the same number of frames to match the experimental setup.

TABLE VII

MOT RESULTS MADE ON THE FRAMES USED FOR ReST METHOD EVALUATION; GARDEN1: 2280-2849, GARDEN2: 4800-6000 AND PARKINGLOT: 5828-6475. IT IS NOT POSSIBLE TO COMPARE METHODS ON WHOLE RECORDINGS BECAUSE THE TEMPORAL AND SPATIAL GRAPHS OF ReST METHOD WERE TRAINED ON PREVIOUS FRAMES.

Method	Metric	Recording		
		Garden1	Garden2	Parkinglot
GRAP-MOT	IDF1	39.84	33.69	38.16
	MOTA	33.11	33.61	23.26
ReST	IDF1	27.3	32.2	25.0
	MOTA	78.5	85.5	76.7

The Auditorium recording was not tested because the authors did not supplement the spatial and temporal graphs of the ReST method. The GRAP-MOT scored best in terms of IDF1 on each recording (Table VII). For Garden1 and Parkinglot the differences in IDF1 values are 12.54 and 13.16 respectively; for Garden2 the difference is 1.49. The ReST method scored best in terms of MOTA, with differences in values being for Garden1 38.66, Garden2 51.86, and 53.44. The reason for such large discrepancies between the IDF1 and MOTA values of the two methods is explained in the discussion, while also showing examples of why IDF1 is a better measure for assessing MOT quality.

IV. DISCUSSION

We proposed a novel method for detecting and tracking people in closed spaces that effectively combines information from multiple video cameras. We tested several solutions for tracking, feature extraction, and community detection on our internal dataset and highlighted the best algorithms: SORT, ResNet50 model, and Greedy Modularity Maximisation community detection method. Then, we used the CAMPUS dataset to compare the GRAP-MOT with other state-of-the-art methods. The requirement of a highly congested space

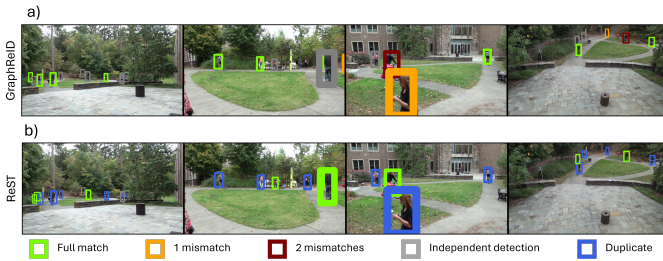


Fig. 4. Example frame from the Garden2 recording illustrating discrepancies between IDF1 and MOTA. (a) GRAP-MOT final pipeline with a few ID mismatches, yielding IDF1 = 38.66 and MOTA = 48.09. (b) ReST with spatial and temporal graphs trained on Garden2 shows low IDF1 = 32.2 but high MOTA = 85.5 due to duplicated IDs. Color coding: green – correctly tracked detections across cameras; orange – mismatch on one camera; red – mismatch on two cameras; blue – repeated ID on one camera; grey – detections missing in other frames.

with overlapping camera views was satisfied by the Garden1, Garden2, and Parkinglot subsets, and the requirement of closed space was met by the Auditorium and Parkinglot subsets. We obtained comparable MOTA metric values to the TRACTA method and better than STP, HCT, KSP, and POM (Table VI). The DyGLIP reports outstanding results of the MOTA parameter, but as we describe later, it's not the best metric. From existing algorithms, we managed to run only the ReST method (Table VII). Large gap between the IDF1 and MOTA when using ReST results mostly from duplicate detections in the scope of one recording. At times when the ReST model can not decide which neighboring detections should have a specific ID, it assigns the same ID to both detections. This eliminates the ID switches, resulting in a larger MOTA, but lower IDF1. An example frame from the Garden2 recording is shown in Figure 4 where the top panel contains GRAP-MOT matches and the bottom panel ReST matches. We did not use the Market1501 [70], Mars [84], or CUHK03 [85] datasets for evaluation since they do not meet the above-specified requirements.

When calculating the Importance metric 6, image bounding box features and position estimation were the base for the MOT anchor, with occurrences solidifying the connection between vertices. When the position estimation module is off 15, only features are a solid base for the MOT anchor; the r bounding box relation coefficient is a rough position estimation meant to strengthen the already existing anchor. This makes the GRAP-MOT approach with position estimation solidify connections between proper vertices much faster because both features and position work together. Without position estimation, the GRAP-MOT first analyzes features, and only after some time checks how bounding boxes are located in relation to each other. The position estimation plays an important role in connecting all the MOT elements; thus, the SORT tracking method reports the best results (Table II). On the other hand, without the position data, deeper information is required, and DeepSORT gives better results (Table V). But still, the difference in medians between the GRAP-MOT with and without the position estimation module for IDF1 is 81.492 to 36.076, and for MOTA 72.269 to 11.076, respectively. Taking those results into consideration,

the position data should improve results for external datasets as well.

Our problem revolves mainly around human multi-camera multi-object tracking. While sharing many similarities with more popular vehicle tracking, it covers a much wider variety of backgrounds and has to deal with occlusion, viewpoint variations, and pose variations. In the person MOT task, aside from the bounding box features, additional information could be used, like trajectory [28], homography matrix information to map detections across cameras [39], pose [59], [60] or specific body part features [86]. However, in the environment presented in our dataset, the use of other body parts features, people's poses, and their walking trajectories is not possible. To supplement this, we demonstrate the effectiveness of the exact position determination method, while also proposing an alternative approach that replaces the use of the homography matrix, acknowledging that it may not always be available. Applying various deep learning methods is the most popular approach [39], [56]–[58] since their ability to gather features from the images is unrivaled. However, the availability of working implementations for person MOT is very limited and, if they exist, it is very difficult to cope with the complexities of the programming environment they impose. Although code for methods such as TRACTA and DyGLIP is available on the GitHub platform, we have not been able to run them. In the case of TRACTA, this was due to requirements such as the CAFFE library, which ceased to be supported in 2014, and its installation requires compiling it from source and installing a suitably old version of Linux under that. In the case of the DyGLIP method, despite our sincere intentions, we noticed that a key script was missing. Without it, the second step (graph features extraction) of the method is impossible to execute. The authors are aware of the absence of this piece of code. In the case of ReST, we managed to run the provided code without major problems.

After careful analysis, we consider IDF1 rather than MOTA the most important metric for person MOT performance testing (but we reported both parameters in our results). In the MOT task, ensuring the correct and consistent assignment of identifiers to detections across multiple cameras is critical for preserving identity tracking. MOTA penalizes the method primarily for errors such as missed detections and false alarms (both directly tied to the presence or absence of the bounding boxes), which means that even incorrectly assigned identifiers could still give high MOTA values. Also, identifier switches are penalized, but a direct measure of the impact of this change on identifier correctness is not provided. In contrast, IDF1 specifically measures the quality of identifier assignments and consistency across all frames. Therefore, we believe that IDF1 is a better scoring metric for MOT tasks than MOTA. There are situations when a high discrepancy between both indices is observed (Supplementary Figure 5). An example of where IDF1 is a better MOT quality measure than MOTA is an experiment concerning community detection methods (Table IV). When analyzing the result of the experiment in terms of MOTA values, the clear winner is GNI (Girvan–Newman weighted by Importance). Post hoc analysis indicates statistical significance over other methods, with values ranging from p-

value=8.389e-07, in the case of GMM, up to p-value=3.4e-38, in the case of SC. This results from stagnation, i.e. about halfway through the recording method stops changing the IDs of the tracklets. This results in no ID switches and a constant rise in MOTA. The label assignment is still wrong, which is neglected by the MOTA metric (Figure 5).

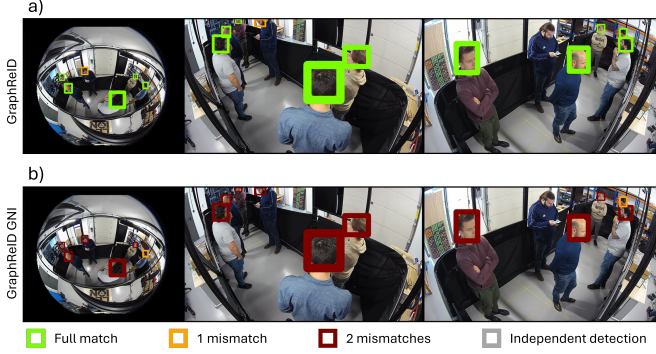


Fig. 5. Example frame from the internal gt6_task10 recording highlighting the mismatch between IDF1 and MOTA. a) GRAP-MOT final pipeline, where both IDF1=97.607 and MOTA=95.214 indicate relatively correct results, b) GRAP-MOT testing GNI community detection algorithm for which mismatches are visible in IDF1=39.567 while MOTA=86.463 is still very high. Green: all detections tracked correctly across the multi-camera views, orange: MOT mismatch on one camera, red: MOT mismatch on two cameras and grey: detections not tracked on other frames.

In a closed space, the main cause of tracklet interference is the presence of obstructions. When an obstruction occurs, a node linked to a disappearing detection stops receiving updates with new image feature values. However, the node itself does not disappear; its stored values remain in memory for a period of time. Community formation continues based on this older information. If a connection is formed during this time, the occurrences value is also updated, which strengthens the connection. This process helps maintain the continuity of the tracklet despite temporary interruptions. To test the robustness of the proposed method against occlusions, we designed an experiment in which occlusions were simulated by removing random detections for 20 consecutive frames every 20 frames from one, two, or three cameras. The tests were performed on 10 recordings from the scene named gt10 and compared to the reference recordings (with all detections present) using a T-test (Table VIII). The T-test shows that there is no statistically significant difference between the IDF1 and MOTA values of recordings with and without occlusions, assuming a significance level $\alpha = 0.05$.

TABLE VIII
MEDIAN IDF1 AND MOTA OF THE GT10 SCENE RECORDINGS, ALONG WITH T-TEST P-VALUES COMPARING THE RECORDINGS WITH OCCLUSIONS SIMULATED ON ONE, TWO, AND THREE CAMERAS TO THE REFERENCE RECORDINGS WITHOUT OCCLUSIONS.

Cameras	IDF1	p-value	MOTA	p-value
-	71.057	-	53.178	-
fe	71.057	1.0	53.178	1.0
fe + left	70.648	0.927	52.826	0.869
fe + left + right	70.410	0.825	52.708	0.863

GRAP-MOT has several limitations. It is designed for

short recordings in enclosed spaces, such as rooms, corridors, warehouses, or public transport, where people are densely packed, partially occluded, and visible from multiple camera angles. In this study, we applied head detection instead of full-body detection to improve visibility and resistance to occlusion. New tracklets are initially uncertain, and their IDs may fluctuate during the first few frames. The method assumes that people within the field of view rarely leave it, and stabilization of tracklet IDs occurs only after the Importance value stabilizes. This delayed stabilization can lower MOTA scores compared to other methods, as seen in the Garden1 recording. Garden1 footage is outdoors, showing a large group of distant individuals who are close together in one camera view. At this distance, feature extraction provides limited information, and proximity-based analysis fails to separate tracklets effectively. Additionally, the frequent entry and exit of people in the early frames prevent the algorithm from achieving stabilization in time.

Future work aims to adapt the system for open spaces without overlapping camera views. Extending the lifespan of graph edges and refining feature management could improve performance in such environments. Network and code optimizations may reduce computational time. The current network is designed for full-body analysis and trained on Market1501 and DukeMTMC datasets, limiting its suitability for the present problem. Retraining on datasets like Wildtrack [87] could enhance efficiency.

V. CONCLUSIONS

In the presented work, we developed a new method for tracking multiple persons under conditions of observation from many cameras in a closed space. We tested the proposed algorithm on publicly available data and on data that we recorded ourselves. We also conducted an in-depth analysis of comparative parameters used to analyze the MOT problem. Finally, we put great emphasis on data access and the transparency, and ease of running our source code.

REFERENCES

- [1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. 1–1.
- [2] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13–es, dec 2006.
- [3] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, p. 61 – 79, 1997.
- [4] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Transactions on Image Processing*, vol. 7, no. 3, p. 359 – 369, 1998.
- [5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, p. 1511–1518.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.
- [7] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, vol. 2015 International Conference on Computer Vision, ICCV 2015, 2015, p. 1440 – 1448.
- [8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, vol. 2015-January, 2015, p. 91 – 99.

- [9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, vol. 2017-October, 2017, p. 2980 – 2988.
- [10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-December, 2016, p. 779 – 788.
- [11] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.
- [12] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, p. 35–45, 1960.
- [13] V. Salari and I. Sethi, "Feature point correspondence in the presence of occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 87–91, 1990.
- [14] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.
- [15] S. Avidan, "Support vector tracking," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–I.
- [16] J. Lasenby, A. Zisserman, R. Cipolla, H. C. Longuet-Higgins, A. Blake, B. Bascle, M. Isard, and J. MacCormick, "Statistical models of visual shape and motion," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 356, no. 1740, pp. 1283–1302, 1998.
- [17] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.
- [18] W. J. Wilson, C. C. W. Hulls, and G. S. Bell, "Relative end-effector control using cartesian position based visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, p. 684 – 696, 1996.
- [19] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, vol. 2016-August, 2016, p. 3464 – 3468.
- [20] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.
- [21] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 748–756.
- [22] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 1–21.
- [23] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. Los Alamitos, CA, USA: IEEE Computer Society, jul 2018, pp. 1–6.
- [24] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, p. 3 – 19, 2013.
- [25] M.-C. Chang, C.-K. Chiang, C.-M. Tsai, Y.-K. Chang, H.-L. Chiang, Y.-A. Wang, S.-Y. Chang, Y.-L. Li, M.-S. Tsai, and H.-Y. Tseng, "Ai city challenge 2020 – computer vision for smart transportation applications," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2638–2647.
- [26] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, Y. Yao, L. Zheng, P. Chakraborty, C. E. Lopez, A. Sharma, Q. Feng, V. Ablavsky, and S. Sclaroff, "The 5th ai city challenge," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.
- [27] E. Luna, J. C. SanMiguel, J. M. Martínez, and M. Escudero-Viñolo, "Online clustering-based multi-camera vehicle tracking in scenarios with overlapping fovs," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 7063–7083, Feb 2022.
- [28] A. Specker, L. Florin, M. Cormier, and J. 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)*, 2022, pp. 3198–3208.
- [29] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2456–2465.
- [30] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann, "Electricity: An efficient multi-camera vehicle tracking system for intelligent city," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2511–2519.
- [31] E. Luna, J. C. S. Miguel, J. M. Martínez, and M. Escudero-Viñolo, "Graph convolutional network for multi-target multi-camera vehicle tracking," 2022.
- [32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [33] H. Hu, R. Hachiuma, H. Saito, Y. Takatsume, and H. Kajita, "Multi-camera multi-person tracking and re-identification in an operating room," *Journal of Imaging*, vol. 8, no. 8, 2022.
- [34] Z. Zhou, D. Yin, J. Ding, Y. Luo, M. Yuan, and C. Zhu, "Collaborative tracking method in multi-camera system," *Journal of Shanghai Jiaotong University (Science)*, vol. 25, no. 6, pp. 802–810, Dec 2020.
- [35] Y. Jeon, D. Q. Tran, M. Park, and S. Park, "Leveraging future trajectory prediction for multi-camera people tracking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 5399–5408.
- [36] Q. You and H. Jiang, "Real-time 3d deep multi-camera tracking," 2020.
- [37] Y. He, X. Wei, X. Hong, W. Shi, and Y. Gong, "Multi-target multi-camera tracking by tracklet-to-target assignment," *IEEE Transactions on Image Processing*, vol. 29, pp. 5191–5205, 2020.
- [38] K. G. Quach, P. Nguyen, H. Le, T.-D. Truong, C. N. Duong, M.-T. Tran, and K. Luu, "Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2021, pp. 13 779–13 788.
- [39] C.-C. Cheng, M.-X. Qiu, C.-K. Chiang, and S.-H. 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 (ICCV)*, October 2023, pp. 10051–10060.
- [40] V. D. Nguyen, S. Mirza, A. Zakeri, A. Gupta, K. Khaldi, R. Aloui, P. Mantini, S. K. Shah, and F. Merchant, "Tackling domain shifts in person re-identification: A survey and analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4149–4159.
- [41] Q. Zhang, L. Wang, V. M. Patel, X. Xie, and J. Lai, "View-decoupled transformer for person re-identification under aerial-ground camera network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 000–22 009.
- [42] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 197–27 206.
- [43] H. Li, M. Ye, M. Zhang, and B. Du, "All in one framework for multi-modal re-identification in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 459–17 469.
- [44] K. Xu, X. Zou, Y. Peng, and J. Zhou, "Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 604–16 613.
- [45] X. Xu, X. Yuan, Z. Wang, K. Zhang, and R. Hu, "Rank-in-rank loss for person re-identification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2s, pp. 1–21, 2022.
- [46] L. Ding, L. Liu, Y. Huang, C. Li, C. Zhang, W. Wang, and L. Wang, "Text-to-image vehicle re-identification: Multi-scale multi-view cross-modal alignment network and a unified benchmark," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 7673–7686, 2024.
- [47] J. Wu, R. Hong, and S. Tang, "Intermediary-generated bridge network for rgb-d cross-modal re-identification," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 6, pp. 1–25, 2024.
- [48] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2265–2274.
- [49] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 486–504.
- [50] H. Li, A. Zheng, L. Sun, and Y. Luo, "Camera topology graph guided vehicle re-identification," *IEEE Transactions on Multimedia*, vol. 26, pp. 1565–1577, 2023.

- [51] L. Bao, B. Ma, H. Chang, and X. Chen, "Masked graph attention network for person re-identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 1496–1505.
- [52] D. Wang, Q. Wang, Z. Tu, W. Min, X. Xiong, Y. Zhong, and D. Gai, "Vision-language constraint graph representation learning for unsupervised vehicle re-identification," *Expert Systems with Applications*, vol. 255, p. 124495, 2024.
- [53] S. U. Khan, I. U. Haq, N. Khan, A. Ullah, K. Muhammad, H. Chen, S. W. Baik, and V. H. C. de Albuquerque, "Efficient person reidentification for iot-assisted cyber-physical systems," *IEEE Internet of Things Journal*, vol. 10, no. 21, p. 18695–18707, Nov. 2023.
- [54] X. Zhang, X. Xie, J. Lai, and W.-S. Zheng, "Cross-camera trajectories help person retrieval in a camera network," *IEEE Transactions on Image Processing*, vol. 32, p. 3806–3820, 2023.
- [55] T. Wang and S.-H. Chiang, "Online pedestrian tracking using a dense fisheye camera network with edge computing," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, Oct. 2023.
- [56] M. Wiecezorek, B. Rychalska, and J. Dabrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *Neural Information Processing*, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds. Cham: Springer International Publishing, 2021, pp. 212–223.
- [57] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen, "Unsupervised pre-training for person re-identification," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 745–14 754.
- [58] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.
- [59] M. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11 2017.
- [60] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "Fd-gan: pose-guided feature distilling gan for robust person re-identification," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 1230–1241.
- [61] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.
- [62] Y. Xu, X. Liu, Y. Liu, and S.-C. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4256–4265.
- [63] Y. Xu, X. Liu, L. Qin, and S.-C. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 4299–4305.
- [64] M. Marczyk, A. Kempinski, M. Socha, M. Cogiel, P. Foszner, and M. Staniszewski, "Passenger location estimation in public transport: Evaluating methods and camera placement impact," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2024.
- [65] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.
- [66] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, sep 2021.
- [67] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," 2019.
- [68] —, "Learning generalisable omni-scale representations for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5056–5069, Sep. 2022.
- [69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [70] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.
- [71] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV Workshops*, 2016.
- [72] K. Zhou and T. Xiang, "Torchreid: A library for deep learning person re-identification in pytorch," *arXiv preprint arXiv:1910.10093*, 2019.
- [73] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [74] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.
- [75] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004.
- [76] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, p. 1–10, 2008.
- [77] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," 2016.
- [78] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 17–35.
- [79] C. Heindl, "py-motmetrics," <https://github.com/cheind/py-motmetrics>, 2023.
- [80] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, oct 2008.
- [81] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E*, vol. 76, p. 036106, Sep 2007.
- [82] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2001.
- [83] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.
- [84] *MARS: A Video Benchmark for Large-Scale Person Re-identification*, Springer. Cham: Springer International Publishing, 2016.
- [85] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.
- [86] V. Somers, C. D. Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person re-identification," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 1613–1623.
- [87] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret, "Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5030–5039.