A Novel Approach for Privacy Preserving Object Re-Identification on Edge Devices

Robert Kathrein
JRZ - Vision2Move
University of Applied Sciences Kufstein
University Passau - Germany
Kufstein, Tyrol, Austria
Robert.Kathrein@fh-kufstein.ac.at

Oliver Zeilerbauer, Johannes Larcher, Mario Döller
JRZ - Vision2Move
University of Applied Sciences Kufstein
Kufstein, Tyrol, Austria
Oliver.Zeilerbauer, Johannes.Larcher, Mario.Doeller@fh-kufstein.ac.at

***Abstract*—Computer vision approaches have been widely used in mobility tasks such as visitor counting, traffic analysis, etc. The European General Data Protection Regulation (GDPR) enforces in-camera processing as storing and transmitting such data violates this regulation.**

This paper introduces a novel approach for object Re-Identification (Re-ID) on edge devices using a color based encoded virtual plane for location mapping. The method leverages the spatial coding capabilities of the RGB color space to simplify the localisation process. By assigning unique RGB values to spatial coordinates, creating a multidimensional reference image that facilitates instant and accurate object localisation. This reduces computational complexity and allows global referencing across multiple cameras. We present an algorithmic framework for location mapping and demonstrating its capability through experimental validation. The techniques potential is further explored in applications such as object Re-ID, marking a significant advancement in computer vision and expanding the branch of spatial encoding methodologies. This approach represents a shift towards more privacy-oriented multi camera object tracking and Re-ID solutions.

I. Introduction

Solving the Re-ID problem in multi camera environments [1] is a well-known, current research field. Especially the topics of object counting, traffic management or behavior analysis requires solid and accurate positional object tracking. This relies on the analysis of trajectories that can be extracted of individual objects. Object Re-Identification (Re-ID) has been defined by assigning a unique ID to the same object (person, car, etc.) across multiple camera systems in an arbitrary environment. There is a differentiation of multi camera setups, setting apart overlapping and non-overlapping camera views. This paper specifically focuses on scenarios where camera systems have overlapping fields of view. The set of overlapping cameras as well as the observed regions can vary. Within this setting, several different techniques and approaches for identifying individuals and/or objects have been proposed [2]. Most of these publications focus on person Re-ID [2] which elaborates on specific features such as pose detection, shape or cloth comparison for solving the matching process. As the observed objects privacy has to be preserved, only selected non-identifying or anonymised features are allowed to be exchanged between camera systems or a centralised server. The usage of highly identifying features like facial features is prohibited in terms of privacy. Besides the extraction of the objects features, every object could potentially be located in the 3D environment during observation. By achieving an accurate object localisation, over time a trajectory can be formed which further helps solving the Re-ID problem. Localisation has been addressed in a few articles [3], [4] so far, but the Re-ID potential of the object position as feature has not been evaluated in detail. Our approach introduces a virtual plane with a color gradient positioned in the virtual representation of the environment, where every camera also has its own digital counterpart. Rendering the virtual environment for each cameras view effectively creates globally correct location reference images. Thus enables to rapidly determine the global position inside the 3D environment of every camera pixel in a singular RGB value. This paper shows the effectiveness of our color based location approach for multiple camera environments, and that solving the Re-ID problem can be supported and solved by exact location information. Furthermore the optimal camera tilt and object rotation for targeting the correct location pixel on the reference image is examined. The work is structured as follows: Section II is dedicated to the different research areas related to this article. Focused on multi object tracking (MOT) in several camera constellations. A detailed description of the developed location reference workflow, including the Re-ID testing setup and datasets is summarised in section III. The results of our accuracy and performance tests of the color based location approach itself as well as the Re-ID testing is summarised in section IV. Finally in section V an extensive conclusion is drawn from all accomplished experiments.

II. Related Work

A. SC-MOT

Single Camera Multi-Object Tracking (SC-MOT) is a key research area in computer vision for tracking multiple objects within a cameras video stream. It involves detecting and maintaining the identities of multiple objects over time, despite challenges like occlusions and changes in appearance. SC-MOT is vital for surveillance, traffic monitoring and activity analysis. The main challenges include accurately detecting

objects in successive frames and associating these detections over time amidst lighting changes, occlusions, and appearance similarities [5], [6]. State-of-the-art systems use advanced deep learning models for object detection and sophisticated data association techniques to maintain consistent tracking [7]. Despite progress, SC-MOT remains a highly active researched area, focusing on improving accuracy, efficiency, and adaptability to various scenarios.

B. MC-MOT

Multi Camera Multi-Object Tracking (MC-MOT) systems are crucial in applications ranging from surveillance to autonomous driving and traffic management, distinguished by **online and offline** tracking methodologies. A further differentiation criteria of MC-MOT systems is the used approach for the Re-ID solution, grading the system in a **Two-Step or Global** type. [8]

1) Online Tracking: processes and analyzes data in near real-time, detecting and tracking objects as video is streamed from cameras. This approach is essential for immediate actions like real-time surveillance and live traffic monitoring. [9] The most frequently used algorithms are SORT [10], DeepSORT [11], ByteTrack [12] and FairMOT [13]. They use only historical information, meaning predictions of object locations in the current frame rely solely on data from previous frames. [7], [8]

2) Offline Tracking: in the other hand, analyzes data post-collection using complex algorithms for high-accuracy scenarios like traffic flow studies. By processing the entire dataset at once, including future points, it makes more informed decisions on object trajectories, enhancing accuracy and handling of complex scenarios. [14], [15]

3) Two-Step: Re-ID approaches are collecting fully associated trajectories of multiple SC-MOT systems. In the second step it uses different association techniques including Bayesian inference [16] or another layer of deep learning methods to achieve a trajectory fusion. [8] This procedure is called Inter-Camera association. The reason to implement such approach is to compare the effectiveness of the MC-MOT system strictly on a new matching algorithm by exert widely used SC-MOT algorithms. [17], [18]

4) Global: Re-ID approaches in contrast are referring to the strategy of tracking objects across a wide network of cameras, across different locations and times with a single unique trajectory per object. This approach digests the features of new object detections, producing new objects or enhancing the trajectories of existing objects dynamically during runtime. [19], [20] A prominent matching algorithm used to combine those detected features with existing objects is the Hungarian Algorithm [21], [22] Achieving effective global tracking requires addressing several challenges [23], [24], such as:

- **Cross-camera feature consistency:** Ensuring that the features used for identifying objects are robust across different camera views and conditions.
- **Spatial-temporal reasoning:** Making sense of the spatial layout of the camera network and the temporal gaps that may occur between sightings of an object in different cameras.
- **Scalability:** Efficiently processing data from potentially large networks of cameras without compromising tracking accuracy or speed.

Global tracking approaches are key in applications such as city-wide surveillance, large-scale event monitoring, and complex security systems, where understanding the movement and behavior of individuals or objects across broad areas is crucial. The approaches presented in [23] and [24], leverage advancements in machine learning, particularly deep learning, to improve feature extraction, matching, and trajectory prediction, thereby enhancing the overall efficacy and reliability of Re-ID systems.

C. Privacy Preserving MC-MOT

The usage of anonymous features is crucial for protecting individual privacy and adhering to ethical standards, especially in applications related to surveillance and personal data processing. As tracking technologies have the capability to track and identify individuals across different spaces and time, they pose significant privacy risks if not managed correctly. In [25] a comprehensive progress is presented where the low level features created by deep learning approaches for object Re-ID can be used to nearly fully reconstruct the initial image. Anonymity is accomplished through various techniques, including pixelation, blurring faces or deep appearance features, and using deep learning models to generate anonymised data that retain essential characteristics for Re-ID purpose-s without revealing actual identities [25].

III. Color Based Location Reference System

This section is dedicated to the novel approach of mapping every pixel of the camera image view to an inter-camera globalized virtual location.

For our Color Based Location Reference System (CBLRS), a MC-MOT environment with partially or fully overlapping camera views is required. The approach is proposed in three distinct steps. The first step focuses on mapping the camera images pixel coordinates to the XY-Position in the location reference environment using a colorized virtual plane. The global location is represented in the RGB color space and mapped via the so called location reference image. Every camera within the MC-MOT system has its own specific reference image which is location synchronised across all cameras. The second step deals with targeting a detected object in such way, that the objects position can be determined by using the generated location reference image of the previous step. The final third step summarizes the actual centralized Re-ID process in which a global approach with an online method is implemented.

A. Mapping location to color space

Let a CBLRS environment be defined as a physical environment observed by a set of cameras $C_{1...n}$ where n is the number of cameras. A camera is defined as $\mathbf{C} = \{\mathbf{x_{pos}}, \mathbf{y_{pos}}, \mathbf{z_{pos}}, \mathbf{x_{rot}}, \mathbf{y_{rot}}, \mathbf{z_{rot}}, \mathbf{I_{original}}, \mathbf{f}, \mathbf{H}, \mathbf{I_{ReID}}\}$. These XYZ-values define the cameras six degrees of freedom (DoF) meaning its exact position (three axis of translation) and orientation (three axis of rotation) forming the extrinsic camera parameters in the 3D environment. Naturally a camera also includes its current image defined as $\mathbf{I_{original_i}}$ where i specifies the camera in the set of cameras. The implied image size w_i as width and h_i as height and additionally the cameras focal length $\mathbf{f_i}$ form the cameras intrinsic parameters. With this information the homography matrix H_i is described. Every parameter till now is used to create a virtual representation (digital twin) of the environment including all cameras. A virtual two dimensional plane p is added the this virtual environment and positioned that the total observed area is covered. By leveraging the capabilities of H_i and applying a texture gradient from 0% to 100% of the color channel red and green on its width p_w and height p_h, this plane can be used to calculate the image of the virtual cameras view. This rendering is defined as reference image formalised as $\mathbf{I_{ReID_i}}$ - completing our camera definition. This I_{ReID_i} is an exact overlay of the $I_{original_i}$ and can be visualized as such.

The generation of such reference image can be accomplished with other approaches as well. For example, by using a game or 3D rendering engine. Such approach can be useful if the exact camera partially parameters are unknown. Recent datasets within the scope of object Re-ID, provide the user with a homography matrix $\mathbf{H}$, which can then be used to transform pixel coordinates into world coordinates of a plane in a 3D space. [26], [27] Let $I_w \in \mathbb{N}$ be the image width and $I_h \in \mathbb{N}$ be the image height, let $p_w \in \mathbb{R}$ be the plane width and $p_l \in \mathbb{R}$ be the plane length. Let $\mathbf{x} = (x_1, y_1, 1)$ with $x_1 \in \{1, 2, ..., I_w\}$ and $y_1 \in \{1, 2, ..., I_h\}$ be the vector describing the pixel-coordinates in homogeneous coordinates and $\mathbf{X} = (X_1, Y_1, 1)$ with $X_1 \in \left(-\frac{p_w}{2}, \frac{p_w}{2}\right)$ and $Y_1 \in \left(-\frac{p_l}{2}, \frac{p_l}{2}\right)$ be the vector describing the worlds-coordinates on the projected plane. Furthermore let $\mathbf{H} \in \mathbb{R}^{3\times3}$ be the corresponding homography-matrix and $s \in \mathbb{R}$ a scaling factor. The following equation allows us to convert pixel-coordinates to worlds-coordinates and back:

$$s * \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \mathbf{H} \times \begin{bmatrix} X_1 \\ Y_1 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11}, h_{12}, h_{13} \\ h_{21}, h_{22}, h_{23} \\ h_{31}, h_{32}, h_{33} \end{bmatrix} \times \begin{bmatrix} X_1 \\ Y_1 \\ 1 \end{bmatrix} \quad (1)$$

For benchmarking purposes, a simulated environment with 15 cameras and their corresponding calculated reference images is created. The calculation of the reference images is shown in figure 2. The extracted color value of our location method can be directly used to calculate pseudo distances between two virtual points. Furthermore, by building the detected objects trajectories over time appending each color point, the resulting lines of color could potentially be compared by using histograms. This investigation would exceed the boundaries of this work. Therefore, this theory will be covered in a future work.

B. Location Targeting

By now, the reference images (I_{ReID}), which can be used to pinpoint exact locations across different cameras are constructed. The research of Re-ID approaches shows, that for the object detection process often an off-the-shelf solution is chosen. The resulting labels are usually rectangular bounding boxes [28]. Other approaches use neuronal networks, producing 3D bounding boxes [29] in the process. But unfortunately, such datasets are rarely available. Naturally an object is raised above the ground XY plane and positioned on a single location point. Yet a bounding box provides a range of pixels, the following question arises: Which of these pixels should be considered as the objects exact point of origin? This consideration gets even more complex when the objects type, its rotation, as well as the downward tilt and angles of the camera are introduced. All of those factors change the dimensions of the bounding box shifting the objects origin.

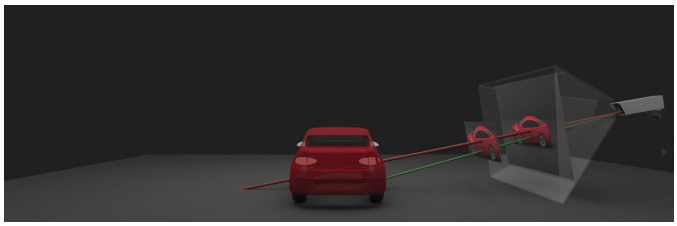

Fig. 1. This images shows a camera detecting a vehicle. When the vertical center point of the label is selected (red ray), the ray would target a position behind the vehicle. The targeting pixel needs to be adjusted depending on camera tilt and object rotation to select the closest possible position pixel of the object. In this example approximately at the green visualised ray.

In a naive approach the center-center point (horizontal-vertical) of the detections label could be considered as the objects location target point. But the XY plane in the reference image is located at ground floor. A ray on the center of the object is cast through the object on a different location on the RGB plane. By targeting the center-bottom point (horizontal-vertical) the selected position is always between the camera and the object but not the real objects origin. This problem is visualised in figure 1. As the authors in [27] described, a lack of datasets with exact object positions is present. By using the Carla (Car Learning to Act) simulation software which is powered by the unreal graphics engine the creation of an synthetic dataset with photo-realistic images is achieved. To tackle this targeting problem with the origin in camera tilt and object rotation, the vehicle simulation framework Carla [30] was used to create a small dataset of camera images from 15 different tilt-angle combinations. Which are presented in figure 2. With this dataset the optimal horizontal target point and object rotation should be determined. The results of this experiment are discussed in section IV.

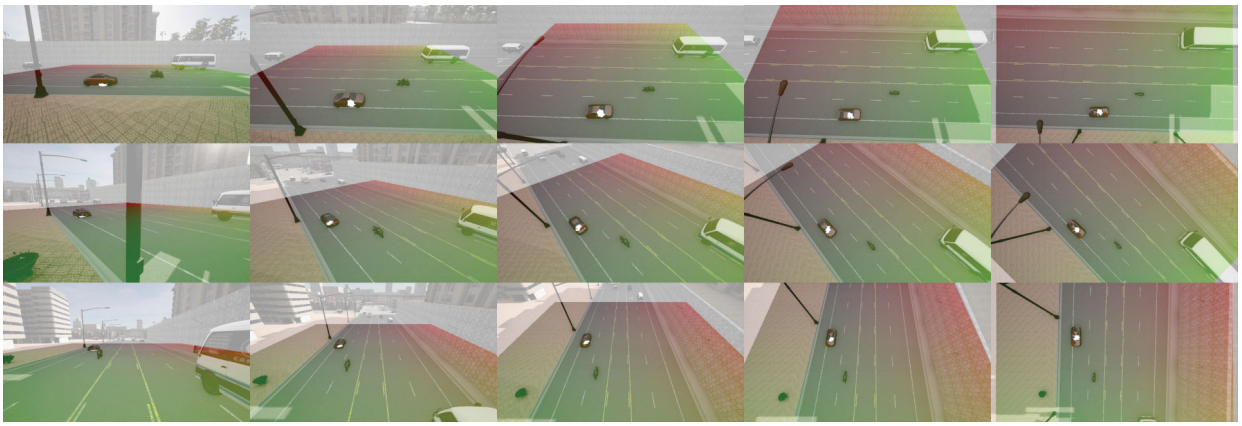

Fig. 2. This figure shows an image collage of 15 cameras with discrete positions and rotations observing an uniform area. The cameras reference image is augmented for better visualisation. As simple showcase a distinct RGB value of the reference image was selected by using the color selection tool of an image software and overwritten with the color white - demonstrating its capability in the process.

C. Basic Re-ID Setup

The colors of these targeted pixels can be used as a pseudo coordinate system for object locations. This simplifies the calculation of distances between detected objects and is enhancing the objects privacy by obfuscating the real location. In section II reasoned methods for MC-MOT are presented typically narrowing to the usage of deep learning algorithms [25]. The result of the presented CBLRS can be used as a new feature set for the object matching process within a Re-ID system. To proof the capability of this feature set, the Carla simulation engine was used to construct a dataset of a dual camera system. The selected area consists of a roundabout with occlusions which is observed by two cameras. A car driving a full circle forces the matching algorithm to hand over the objects ID between the cameras twice. Figure 3 shows the view of both cameras with the overlaid reference image. In total, two small datasets, with two time synchronised videos of 15 and 30 seconds length and 20 FPS were created. Furthermore a basic Re-ID processing pipeline was created with online tracking and a global Re-ID solution approach.

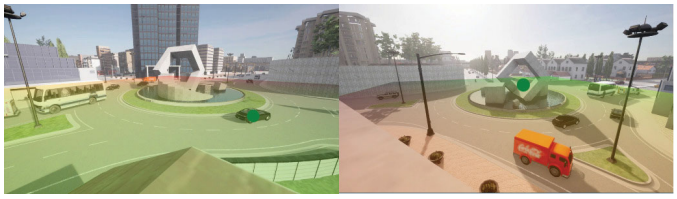

Fig. 3. Visualisation of the camera images including the overlaid reference image for object localisation. The green dot visualises a distinct color value of the reference images. Displaying the capability of matching positions across multiple cameras.

1) Camera Pipeline:: The mentioned pipeline consists of two parts and are visualized in figure 4. The first one is processed directly within each camera. The process starts by capturing the latest camera frame $I_{original_i}$ where i defines the timestamp followed by a generic off-the-shelf labeling method. In our case a pretrained YOLOv5 nano network was trained with approximately 4000 traffic images in various scenarios (real and synthetic). This process creates a set of detections per camera frame defined as $D_i = \{d_1, d_2, ...\}$. A detection is defined by its type t which is derived from the labeling methods class list and its bounding box extends resulting in a this definition $d_j = \{t, x_{left}, x_{right}, y_{top}, y_{bottom}\}$ for $1 \leq j \leq n$ and $n \in \mathbb{N}$. After the detection step, our CBLRS is applied on on every detection using I_{ReID} to compute the color based location information l defined as $l = \{r, g, b\}$ is attach to each detection d_j. As final step the collected detections with their corresponding location data, in combination with the frame number or timestamp of capturing, are transmitted to the central processing node by a message M. Such message is defined as $M = \{i, D_i\}$. Summarised the 4 steps of the detection part of the pipeline are: capturing, labeling, location and transmitting.

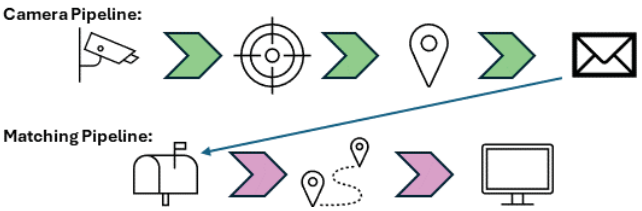

Fig. 4. This info graphic displays processes and connection between the two implemented pipelines. Every individual camera houses the camera pipeline, consisting in capturing, labeling, locating and transmitting data. A global matching pipeline is receiving such camera data, builds the trajectories by matching data and provides visualisation tools.

2) Matching Pipeline:: The second part of the pipeline is located within the central processing system. This part collects and combines all messages received every cameras within the system. Every past detection is collected, stored in an objects data field and used to compare against up coming detections.

At time t we are given an image of the current frame $I^{(t)} \in \mathbb{R}^{W \times H \times 3}$ where W is the image width and H is the image height as well as the current detections $D = d_1, d_2, ..., d_n$ where n is the number of detections. Each

detection $d = (u, c^T)$ contains a timestamp $u \in \mathbb{R}$ and a RGB value $c \in \mathbb{R}^3$ which resembles the transformed bounding-box-coordinates on the RGB plane as explained in equation 1. An arbitrary object $O_i = \{a_1, a_2, ...\}$, where a is an associated detection from a past image $I^{(s)}$ for $1 \leq s \leq t$ which contains a timestamp $u \in \mathbb{R}$ and a RGB value $c \in \mathbb{R}^3$ from a detection as well as the ID $i \in \mathbb{N}$ from the object O_i. The current position of an object O_j is defined as the RGB value of its latest associated detection a. As soon as a message with a new set of detections D is received the euclidean distance between every detection and every object within the system is calculated.

This distance is used as cost parameter for a Hungarian matching algorithm. If a match exceeds a certain threshold in distance, a new object is created. If not, the detection gets appended to the closest corresponding object. Summarised, the three steps of the matching part of the pipeline are: receiving, matching and visualising.

IV. Evaluation

The following section presents a comprehensive analysis of the conducted tests designed to measure the CBLRS's effectiveness and reliability. At first the datasets to conduct the evaluations with where selected. The following tests include performance benchmarks, targeting accuracy, and Re-ID capability. These assessments were meticulously chosen to provide a holistic view of the system's capabilities in various operational scenarios. Performance benchmarks evaluates the system's speed and efficiency compared to other techniques. Targeting accuracy examines the optimal camera orientations for locating targets. And Re-ID capability examines the ability of location information to correctly link related objects and compares our CBLRS approach with a lookup table approach. Together, these tests offer valuable insights into the system's strengths and areas for improvement.

A. Dataset Selection

Selecting an appropriate dataset for MC-MOT tasks can be challenging, particularly when the dataset does not include homography matrices. As show in section III, homography matrices are useful in tracking scenarios for mapping object positions between different camera views or correcting perspective distortions. [31] Especially for our proposed method a homography matrix is mandatory. However, many public MOT datasets do not provide these matrices, necessitating alternative strategies. The only MOT dataset which provides this information is the *Synthehicle* dataset [27]. This dataset consists of different traffic scenes of mostly non overlapping multiple camera views. As the name suggests all data is synthetically generated by using the Carla Car Simulation software [30]. As comparison metric the authors of Synthehicle provide a Multi Object Tracking Accuracy (MOTA) score. Our work is focused on maximising the localisation accuracy of synchronised camera images. Therefor we need to compare our work with a Higher Order Tracking Accuracy (HOTA) score as MOTA does not incorporate localisation accuracy. [32] Carla was used to create a new dataset with overlapping camera views suited to validate our initial workflow.

During the evaluating our results, the CityAI Challenge 2024 released a of dataset. [33] Exposing a brand new MOT dataset of overlapping camera views including homography matrices for all cameras. Furthermore a validation script to calculate the HOTA score (consisting of detection, association and localisation accuracy) is included. An example of camera views are visualised in figure 5. The dataset consists of multiple scenes of warehouse environments where synthetic people have to be tracked. The method proposed in this work was used to tackle this challenge.

B. Object Targeting Accuracy

The vehicles in the self-created dataset are visualised in figure 2. They are rotated in three different angles (top to bottom) of 0°, 45°and 90°combined with five different camera tilts (left to right) of 10°, 30°, 50°, 70°and 90°. This dataset was labeled by hand with bounding boxes and unique vehicle IDs. The Carla simulation software allows to extract the object position what is used as ground truth (GT). For position targeting the center of the labels horizontal axis was chosen. The vertical axis was split in 5% steps from center (50%) till bottom (95%). Every targeted position pixel was compared to the GT by calculating the euclidean distance between the targeted pixel and GT.

Since the Carla simulation provides the coordinates in meters the distance from target to GT in meters or centimeters was selected as metric. The results show the average offset distance (AOD) to the estimated object origin is 112cm. Depending on the camera tilt, the best result was achieved at 30°with an AOD of 56cm to the vehicles origin, peaking at a vertical targeting position (VTP) of 65% with an AOD of 30cm. Comparing the objects rotation, the best result was achieved, when targeting 45°rotated objects with an AOD of 74cm. The full result will be published after acceptance, and submitted to supplementary files peaking at a VTP of 55% with an AOD of 53cm.

C. Performance Benchmark

For every detection the homography can be used to directly calculate the objects position. This calculation involves multiple matrix calculations, therefore the complexity is naturally high. The most potent algorithm, the Strassen algorithm is used to perform such calculations with a complexity of $\mathcal{O}(n^{2.81})$. [34] Furthermore camera systems used for traffic monitoring typically have limited processing power which is better used for labeling the image data. To outsource this calculation the raw images need to be transmitted via a network. But the transmission of raw images can be conflicting with the EU general data protection regulation (GDPR). [35] To preserve the privacy of detected objects and still sustain a high frame rate count for labeling, the color based mapping is introduced.

Our CBLRS utilises a color gradient on a virtual plane to create a reference between pixel coordinates and object position. It was tested using 8 bit color space, where total

Fig. 5. A sample data frame of the CityAI challenge dataset.

of 65.536 distinct locations can be addressed on a single plane, which refers to 16 bit. In a theoretical scenario where a roundabout is observed and the virtual XY plane extends for 50x50m a resolution of 19cm per pixel can be achieved in both axes. Depending on the plane size and required accuracy the channel depth can get extended from their original 8 bit up to 16 bit or higher, doubling the accuracy for each additional bit. Furthermore the by now unused blue color channel can be leveraged to position multiple virtual planes for e.g. a city wide spanning MC-MOT systems. The blue channel can then be used to identify a specific plane. Raising the distinct locations to 16.777.216 when using 255 planes. As these reference images are generated beforehand, the complexity during runtime for a precise localisation is reduced to $\mathcal{O}(1)$. The location representation in RGB values furthermore allows for direct distance calculations as well as obfuscating the objects real coordinates thus boosting the privacy factor.

In real world examples, such cameras are often powered by small edge devices, therefor we conducted a benchmark on processing speed of several edge devices, including the Jetson Nano, Jetson Xavier, Raspberry Pi 3, and Raspberry Pi 4. The performance was compared between the different location processing methods: the homography calculation itself, usage of a lookup table (LUT) and our color based model. Our tests revealed significant variations in processing efficiency across these devices.

The results are displayed in table IV-B. The Jetson Xavier outperformed the others due to its superior computational power, handling all methods with ease, particularly excelling in homography calculation despite its complexity. The Jetson Nano also performed well in LUT and RGB image processing, though it lagged significantly in homography tasks. The Raspberry Pi 4 showed respectable performance, significantly improving over the Raspberry Pi 3, which struggled with the more computation-intensive homography calculations. Overall, the Jetson Xavier demonstrated the best all-around performance, while the Raspberry Pi 4 offered a balanced and cost-effective solution for less demanding tasks. Across all devices the difference between those methods is consistent. LUT compared to calculation of homography matrices is 3 to 5 times faster. Whereas accessing a RGB pixel compared to LUT is approximately $20 - 25\%$ faster. In terms of file storage a singular LUT file for a 1080p image is 31.6 MiB in size whereas the equivalent color coded reference image in PNG file format only takes 143.4 KiB disk space. The JSON file containing the homography matrix only uses about 500 Bytes.

D. MC-MOT / Re-ID Benchmark

In the following page we discuss the used visualisation method for subjective result examination on our Re-ID implementation. Afterwards our findings on the two datasets (Carla, CityAI24) are presented and discussed in detail.

1) Data visualisation:: During processing, the objects are enriched with their temporal spacial information, forming their trajectory. These trajectories are visualised and mapped to a birds eye image of the observed area. Figure 6 shows such visualisation.

2) Carla Dataset:: After every frame has been processed the resulting objects with their associated predictions were examined. In total 606 messages with a total of 2410 detections were processed. All of those are combined and matched to a total number of 19 vehicles. The ground truth has 21 registered

TABLE I. BENCHMARK OF DIFFERENT LOCALISATION METHODS OF DIFFERENT EDGE DEVICES. EACH METHOD WAS RAN 100.000 TIMES (N = 100000), ACCESSING A RANDOM CAMERA PIXEL. THE RESULT VALUES REPRESENT THE ECLIPSED TIME IN SECONDS PROCESSING TOOK. THE RANDOMNESS WAS UNIFIED BY USING A SET RANDOM SEED.

n=100000	i5-4690K @ 3.50GHz	Pi 3 @ 1.2GHz	Pi 4 @ 2GHz	Jetson Nano @ 1.6GHz	Jetson Xavier @ 2.2GHz
Homography	0.84073	30.7385	19.16133	115.0698	4.98495
Lookup Table	0.24109	10.2173	5.80970	31.0034	1.07157
Color based (ours)	0.19545	8.40112	4.59746	23.1562	0.81251

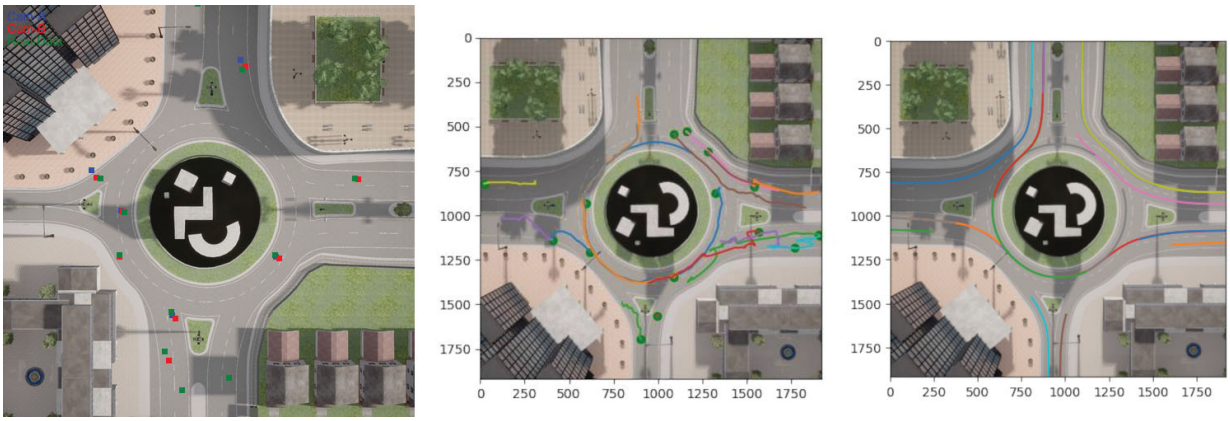

Fig. 6. The left image shows the predictions of a single frame visualised against the GT data, the middle image shows the predicted trajectories and the right image displays the ground truth.

vehicles in the observed area. One of the vehicles was not recognised, because our YOLO was not trained to detect motorcycles. The other undetected object is not yet entered the Field of View (FOV) of any camera. The average distance from each prediction to a object origin is 1.8 meters. The figure 6 visualises the results of this experiment.

3) CityAI 2024 Dataset:: Compared to the self created dataset, the camera angles of this dataset are steeper and the area to observe is smaller. Therefor the RGB reference area was condensed to a square with edges of 30m centered around the world center. This results in a localisation raster of 12x12cm. We conducted our tests on the validation scene 43 of the dataset. A YOLOv8 medium model was trained on the datasets training data (scenes 1-40). Scene 43 is composed of 10 camera videos with in approximately 240.000 frames to process in total with about 6 detected objects per frame resulting in a total of 1.454.348 detected objects. The CityAI dataset provides us with a method to evaluate our results, giving us the metric of Higher Order of Tracking Accuracy (HOTA) [32]. Fundamentally HOTA is calculated by combining the systems performance in detecting (Detection Accuracy - DetA), associating detections to the correct object (Association Accuracy - AssA) and assign a location (Location Accuracy - LocA). The full extend of this metric is not in the scope of this paper but can be researched in this work [32].

TABLE II. VALIDATION RESULT OF CITYAI DATASET SCENE 43.

	Lookup-Table	Ours
HOTA	33.711	32.903
DetA	86.667	85.924
AssA	13.126	12.614
LocA	94.596	93.857

As our Re-ID approach purely uses the detected objects location as association feature a weak result of association accuracy (AssA) was expected. With the RGB approach we were able to archive a Location Accuracy (LocA) value of 93.86% and a total HOTA of 32.90%. For comparison, we exchanged the localisation approach to the LUT implementation of the performance benchmark. The results show, that a ”per-camera” localisation speed increase from approximately 20% decreases the pipelines LocA about 0.74 percentage points (0.8%) and the HOTA metric about 0.8 percentage points HOTA (2.4%), showing that by limiting location resolution the processing speed can be enhanced.

V. CONCLUSION

Recent research in MC-MOT has increasingly focused on deep learning methods to address the data association problem in Re-ID. This shift has been driven by the inadequacy of traditional algorithms to effectively calculate distances between feature sets. In this work, we introduce our novel spatial feature to MC-MOT and Re-ID, demonstrating significant potential and jet fully preserve the privacy of tracked identities. As the only used feature is the entities location and this feature is defined in a virtual pseudo environment. Our proposed methods enhance the precision of label targeting but also open new avenues for integrating prediction-based methods into MC-MOT algorithms. This advancement allows for more accurate tracking of both the precise labels and the global movement of objects.

During the course of our research, we found that the existing literature did not provide any datasets with the necessary camera positions and information required to calculate reference images. Consequently, we created a small dataset to demonstrate the effectiveness of our proposed feature. The AI-City Challenge 2024 released a synthetic dataset that, for the first time, included the homography matrix of the capturing device. This development provided an opportunity for an active comparison of our approach using the newly available dataset.

This work introduces a novel approach to enhance object localization performance on edge devices with minimal impact on accuracy. The method’s elegance lies in its simplicity and visualization capabilities. Additionally, the proposed pipeline can serve as a baseline for developing modular feature extraction methods to further enhance Re-ID potential. The sophisticated reasoning and advancements achieved in two-dimensional location targeting (Section III-B) demonstrate

significant research potential for further improvements in this specialized area.

ACKNOWLEDGMENT

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

REFERENCES

[1] N. Peri, P. Khorramshahi, S. S. Rambhatla, V. Shenoy, S. Rawat, J.-C. Chen, and R. Chellappa, "Towards Real-Time Systems for Vehicle Re-Identification, Multi-Camera Tracking, and Anomaly Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 2648–2657. [Online]. Available: https://ieeexplore.ieee.org/document/9150744/

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," Jan. 2021, arXiv:2001.04193 [cs]. [Online]. Available: http://arxiv.org/abs/2001.04193

[3] 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. [Online]. Available: https://ieeexplore.ieee.org/document/9042858/

[4] W. Sun, X. Chen, X. Zhang, G. Dai, P. Chang, and X. He, "A Multi-Feature Learning Model with Enhanced Local Attention for Vehicle Re-Identification," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3549–3561, 2021. [Online]. Available: https://www.techscience.com/cmc/v69n3/44189

[5] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, pp. 3823–3831, 11 2011.

[6] C. M, G. R, and V. T, "Multiple objects tracking in surveillance video using color and hu moments," *Signal and Image Processing: An International Journal*, vol. 7, no. 3, p. 15–27, Jun. 2016. [Online]. Available: http://dx.doi.org/10.5121/sipij.2016.7302

[7] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," Mar. 2017, arXiv:1703.07402 [cs]. [Online]. Available: http://arxiv.org/abs/1703.07402

[8] T. I. Amosa, P. Sebastian, L. I. Izhar, O. Ibrahim, L. S. Ayinla, A. A. Bahashwan, A. Bala, and Y. A. Samaila, "Multi-camera multi-object tracking: A review of current trends and future advances," *Neurocomputing*, vol. 552, p. 126558, Oct. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231223006811

[9] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura, "FCN-rLSTM: Deep Spatio-Temporal Neural Networks for Vehicle Counting in City Cameras," Jul. 2017, arXiv:1707.09476 [cs]. [Online]. Available: http://arxiv.org/abs/1707.09476

[10] 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)*. IEEE, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2016.7533003

[11] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017.

[12] 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," 2022.

[13] 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, p. 3069–3087, Sep. 2021. [Online]. Available: http://dx.doi.org/10.1007/s11263-021-01513-4

[14] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple Object Tracking: A Literature Review," *Artificial Intelligence*, vol. 293, p. 103448, Apr. 2021, arXiv:1409.7618 [cs]. [Online]. Available: http://arxiv.org/abs/1409.7618

[15] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," May 2016, arXiv:1603.00831 [cs]. [Online]. Available: http://arxiv.org/abs/1603.00831

[16] J. Black, T. Ellis, and P. Rosin, "Multi view image surveillance and tracking," in *Multi view image surveillance and tracking*, 01 2003, pp. 169– 174.

[17] X. Zhou, V. Koltun, and P. Kraehenbuehl, *Tracking Objects as Points*. Springer International Publishing, 2020, p. 474–490.

[18] 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," 2021.

[19] W. Chen, L. Cao, X. Chen, and K. Huang, "An equalised global graphical model-based approach for multi-camera object tracking," 2016.

[20] A. Dehghan, H. Idrees, A. Roshan Zamir, and M. Shah, "Keynote: Automatic detection and tracking of pedestrians in videos with various crowd densities," in *In Proceedings of PED*, 2012.

[21] S.-H. Bae and K.-J. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 595–610, 2018.

[22] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1918–1925.

[23] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking," Sep. 2016, arXiv:1609.01775 [cs]. [Online]. Available: http://arxiv.org/abs/1609.01775

[24] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," Oct. 2016, arXiv:1610.02984 [cs]. [Online]. Available: http://arxiv.org/abs/1610.02984

[25] D. Dangwal, V. T. Lee, H. J. Kim, T. Shen, M. Cowan, R. Shah, C. Trippel, B. Reagen, T. Sherwood, V. Balntas, A. Alaghi, and E. Ilg, "Analysis and mitigations of reverse engineering attacks on local feature descriptors," 2021.

[26] S. Wang, D. C. Anastasiu, Z. Tang, and e. a. Ming-Ching Chang, "The 8th ai city challenge," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024.

[27] F. Herzog, J. Chen, T. Teepe, J. Gilg, S. Hörmann, and G. Rigoll, "Synthehicle: Multi-vehicle multi-camera tracking in virtual cities." [Online]. Available: http://arxiv.org/abs/2208.14167

[28] 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, Nov. 2021, arXiv:2004.01888 [cs]. [Online]. Available: http://arxiv.org/abs/2004.01888

[29] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," 2017. [Online]. Available: https://arxiv.org/abs/1612.00496

[30] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[31] Q. Luong and O. Faugeras, "Determining the fundamental matrix with planes: instability and new algorithms," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 489–494, 1993.

[32] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixe, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *International Journal of Computer Vision*, vol. 129, no. 2, p. 548–578, Oct. 2020.

[33] S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, P. Chakraborty, S. Prajapati, Q. Kong, N. Kobori, M. Gochoo, M.-E. Otgonbold, G. Batnasan, F. Alnajjar, P.-Y. Chen, J.-W. Hsieh, X. Wu, S. S. Pusegaonkar, Y. Wang, S. Biswas, and R. Chellappa, "The 8th AI City Challenge," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024.

[34] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000. [Online]. Available: http://ieeexplore.ieee.org/document/888718/

[35] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council," in *Regulation 2016/679*, 2023. [Online]. Available: https://gdpr-info.eu/art-5-gdpr/