Comprehensive Framework for Cyclist and Helmet Detection, Tracking, and Re-identification in Urban Environments

NHL Stenden Lectoraat in Computer Vision & Data Science

Maikel Boonstra

Supervisors: Maya Aghaei Gavari, Klaas Dijkstra

Abstract—The municipality of Friesland in the Netherlands is researching for a method to quantify the volume of passing cyclists and determine the proportion of people wearing helmets in response to the rising number of accidents resulting in severe injuries. The detection and monitoring of cyclists and helmets in public spaces is the main focus of this study. To do this, we introduce our own dataset made up of three cameras with overlapping field of view, a tracker using a Kalman filter and a Hungarian algorithm, and a Re-identification study to create feature embedding and compare them to match them with a embedding from a detection in a different camera perspective. Our findings reveal that cyclist detection is most accurately accomplished from a side view. Moreover, the application of Re-identification between multiple cameras significantly improves the overall performance of the model, enabling more precise predictions of the number of cyclists and helmets present within the dataset.

Index Terms—[Object detection, Tracking, RE-ID, Cyclists, Helmets, Yolov5]

1 Introduction

The Netherlands has long been known for its love of cycling, with its citizens traveling the greatest distance and making the most trips by bicycle compared to other countries. However, the rise of the e-bike has had a significant impact on transportation in the country. In addition to reducing the use of conventional bicycles, e-bikes have also led to a decrease in car trips and even affected the use of public transportation [1].

While the use high degree of cyclists has had many positive effects on transportation in the Netherlands, there have also been negative consequences. One of these is a recent increase in fatal injuries among cyclists. According to a national investigation [2], the number of fatal bicycle accidents has increased by about 10% since 2015, with people over 50 being the most affected. Out of the 170 fatalities that occurred last year, 60 included collisions with automobiles, 42 with fixed objects like buses or trucks, and 68 involved no collisions at all [3].

In a study by [4], an lower association was found between traumatic head injuries and individuals who frequently wore helmets, as opposed to those who did not typically wear helmets. Despite research showing that proper helmet use can significantly reduce the risk and severity of head injuries in accidents, helmet adoption rates have remained low. While helmets do not prevent accidents, they can mitigate the effects of an accident. According to [5], it is advised to wear a helmet in order to decrease the quantity of fatalities and traumatic brain injuries caused by bicycle accidents.

Monitoring the helmet usage in traffic can be done in different ways. A simple way is to get a team to stand next to the road and do the counting manually. This is an expensive andt time-consuming approach. Alternatively deep learning can be used to recognise cyclists and helmets. Previous research in this topic has been done on detecting helmet usage among motorcycles [6]. Our particular interest is whether a certain perspective favours the tracking and detection of cyclists and helmets. Additionally we want to apply Re-identification between the objects seen from various cameras with the aim to improve the certainty of a cyclist wearing an helmet.

In this paper, our focus is on developing a deep learning and computer vision system to count bicyclists and determine whether or not they are wearing helmets. In order to do that we will need to combine the cyclists in different frames together and detect the cyclist and helmets in the frames. Doing this for each camera perspective enables us to make a comparison between the perspectives. Finally, we aim to combine data from multiple camera views to improve tracking accuracy. As a result, the following research question will be addressed:

- What is the performance of object detection and tracking on cyclists and helmets seen from different perspectives?
- How can Re-identification help in on the counting of cyclists and helmets in the overlapping view using a multitude of cameras?

This is a complex and challenging task, but one that has significant potential for improving bicycle safety. By answering this research question, we aim to contribute by developing a machine learning and computer vision system for multi object multi camera tracking of bicyclists which until so far has not been tried.

1.1 Related Work

1.1.1 Motorhelmet detection

The detection of motorcycle helmets is a problem similar to that of bicycle helmet detection, where the goal is to determine whether or not the rider is wearing a helmet. Traditional methods typically employ a two-stage approach, where the first step is to detect the motorcycle and rider itself from the background. This can be achieved using background subtraction or some form of segmentation with techniques such as Gaussian Mixture Model [7]. Another approach uses classifiers and feature descriptors, such as histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT), and local binary patterns [8]. Recent research has explored the use of wavelet transform combined with random forest for motorcycle detection [9]. Even though this research had very good results on the dataset they used the biggest limitation was the accuracy of the background, where the background subtraction took a long time and could only find moving objects.

- *Maikel Boonstra is a Computer Vision Data Science student at the NHL Stenden University of Applied Sciences, E-mail: maikel.boonstra@student.nhlstenden.com.*
- *Klaas Dijkstra is a lector at the NHL Stenden Lectoraat in Computer Vision & Data Science, E-mail: klaas.dijkstra@nhlstenden.com.*
- *Maya Aghaei Gavari is a researcher at the NHL Stenden Lectoraat in Computer Vision & Data Science, E-mail: maya.aghaei.gavari@nhlstenden.com.*

In the second stage, once the rider was detected, [10] made use of a Convolutional Neural Network (CNN) on the upper one fourth part of the detection to determine if the rider wore a helmet, their reasoning was that helmets should be worn on top of their heads and thus only have to try to detect the helmets there. Opposed to the previously mentioned studies where they used classifiers such as Support Vector Machines or K-Nearest Neighbor [11] to detect whether the rider and/or passengers were wearing a helmet or not.

The aforementioned studies have shown the potential of machine learning approaches to recognizing motorcycles and helmets, suggesting new ways to recognize riders and whether they are wearing a helmet. We seek to extend the well-researched field of helmet detection to bicycle helmets. Therefore we resorted to similar problems to find effective approaches. However there remains a need for further advancements in terms of accuracy and speed [12], particularly when dealing with challenging environments characterized by complex backgrounds or occlusions. Detecting motorcycle helmets, in particular, poses a significant challenge that necessitates the development of robust and efficient techniques.

1.1.2 Tracking

In the realm of Multiple Object Tracking (MOT), the tracking of pedestrians is most common [11]. For a number of reasons MOT of pedestrians is useful for a lot of practical applications, such as cyclist tracking.Pedestrian tracking is interesting because they are non-rigid object and have a lot of intra-class variation. Recently, MOT has a lot of traction and can be implemented in various ways [13].

An architecture (MCMOT) was devised by a Hou. et al. [12] to combine the inputs from several synchronized cameras with their intrinsic and extrinsic properties to determine the target location. Using this approximated location they apply a greedy heuristic to connect the locations of the different cameras into one. To keep track of the targets Hou. et al. use a k-shortest path algorithm. Alternatively another research, [14] showed a approach (MVDet) based on Feature Perspective Transformation. Instead of making an anchor box or point, they concatenate the projection of the convolution feature maps via perspective transformations and use a trained CNN on the ground plane feature map to extract target locations.

A more recent approach got introduced by [15]. They make use of tracklets, which are short series of tracked objects. These contain information such as appearance coherency and motion properties which are used to improve the predictions on the objects. Other approaches are making use off Re-identification based on different levels of features in the CNN [16]. Using this technique it is possible to match people based on their appearances using the feature space. Which was successfully except in the cases where people made sudden turn or changed their paths when temporary occluded. The evidence presented in this section suggests that tracking of people is possible by mapping them on the ground floor and recent studies have used features of persons to track their movements.

Alternatively Simple Online Realtime Tracking (SORT) [17] got introduced. Which is a tracking algorithm that only makes use of a Kalman filter [18] and an hungarian algoritm [19] to track an object. An improved on SORT was made, called Deep-SORT [20] through the addition of a cascading association step that makes use of CNN-based object appearance features. The Mahalanobis distance between object states and the similarity of the object appearance features are combined in the data association algorithm, which later uses the SORT's data association for unmatched states. Even with this addition the frame rate was promising on the object tracking benchmarks.

Although the state-of-the-art techniques discussed in this section have made notable advancements in helmet detection, motorcycle detection, and pedestrian tracking, their reliance on a two-stage approach often leads to time-consuming inference. To address this limitation, we propose the utilization of YOLOv5 [21], a deep learning architecture known for its high accuracy and efficient processing speed. By adopting YOLOv5 for helmet detection, we aim to improve the real-time inference capability of our system.

Furthermore, we propose the implementation of a variation of DeepSORT called Norfair [22]. Norfair uses the same algorithms as DeepSORT with a simple implementation. We anticipate that this tracking algorithm will provide the necessary capabilities to achieve accurate and efficient tracking of cyclists in real-time scenarios.

Overall, by combining the speed and accuracy of YOLOv5 for helmet detection with Norfair, our proposed approach aims to overcome the limitations of the traditional two-stage methods and determine which angle gives the most best tracking.

2 Materials and Methods

The achievement of the research objectives involved a series of systematic actions. Initially, a deep learning approach was employed to detect specific objects within the video frames. This step aimed to accurately identify and locate the objects of interest, specifically cyclists and helmets.

Subsequently, an object tracking technology was utilized to generate a continuous trajectory of bounding boxes for each cyclist across multiple frames. This tracking technology enabled ongoing monitoring of the cyclists movement throughout the video sequence.

Furthermore, a comprehensive dataset comprising various locations and multiple cameras with overlapping views was utilized to capture extensive information about the objects. Individual models were trained for each camera to optimize object detection performance, facilitating the models' ability to learn from diverse environments.

Additionally, Re-identification techniques were applied to integrate information obtained from one camera with data acquired from another camera, with a shared focus on the same object. This integration aimed to enhance the accuracy and reliability of object Re-identification across multiple camera views, thereby improving the overall tracking performance. See also figure 1

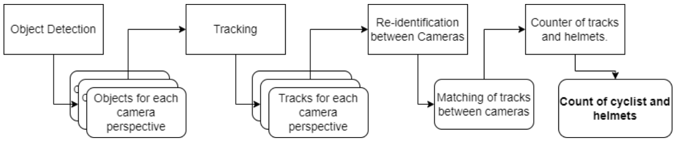

Object Detection
Objects for each camera perspective
Tracking
Tracks for each camera perspective
Re-identification between Cameras
Matching of tracks between cameras
Counter of tracks and helmets.
Count of cyclist and helmets

Fig. 1: This diagram shows the steps our algorithm takes to get a count of cyclists and helmets

2.1 datasets

To facilitate the detection and tracking of cyclists wearing helmets, three datasets were acquired. The first dataset acquisition site (hereafter referred to as the bikeroad dataset) was chosen near the main entry from nearby villages into a city, enabling for the recording of a large number of cyclists. The second location (hereafter referred to as the roundabout dataset) was captured at an intersection of bicycle lanes near a busy roundabout. However, a significant imbalance was identified in both groups within this dataset, as the occurrence of helmet-wearing cyclists was quite low. To address this imbalance, a small group of volunteers rode past the acquisition location while wearing helmets, allowing us to add more cyclists wearing helmets to the dataset. Nonetheless, it is critical to recognize the risk of over-fitting to this specific subset of helmet-wearing cyclists. To mitigate that a 3rd dataset (hereafter referred to as the campus dataset) was acquired. This dataset has a higher percentage of cyclists and helmets while having a smaller variety in cyclists.

The three datasets, to be seen in 3 utilized in this study were partitioned into separate sets, namely a training set, validation set, and testing set, following a split of 70%, 20%, and 10% respectively. This split is quite normal in deep learning, where a large part is delegated for training and a smaller part for validation. The test subset will not be used in this research but is included to give a full overview of the dataset. These datasets were annotated with bounding box annotations for two distinct classes: cyclists and helmets. It is

Table 1: the amount of images, cyclist objects and helmets in the different datasets

Locations		images	helmets	cyclists
Bikeroad	train	6300	184	937
	val	1800	117	347
	test	900	62	498
	total	9000	363	1782
Roundabout	train	2660	765	1651
	val	760	179	739
	test	380	272	1214
	total	3800	1216	3604
Campus	train	1960	898	2102
	val	560	247	594
	test	280	176	476
	total	2800	1321	3172
Campus testing set		images	helmets	cyclists
	Cam1	1494	874	2197
	Cam2	1494	840	2972
	Cam3	1494	844	1334

noteworthy that all datasets were acquired under consistent lighting conditions, specifically during the early morning, characterized by favorable weather conditions devoid of rain, fog, or clouds. Due to ethical considerations, precautions were taken to ensure compliance with the General Data Protection Regulation [23]. Meaning that this dataset is not open to the public and cannot be shared.

The cyclist and helmets annotations are bounding boxes in all of the aforementioned datasets, in order to evaluate the tracking capability of our proposed model we also need to include a tracking ground truth and consecutive images. Therefore we another acquisition was made at the campus locations (hereafter referred to as the campus testset) which we annotated with the included tracking id. This aquisition captures 10 cyclists that drive through the dataset of which 6 are wearing helmets.

We produced these three datasets in an effort to create a dataset with a balanced amount of cyclists and helmets. The number of images, helmets and cyclists in the images can be seen in Table 1. The campus dataset offers a higher percentage of helmet-wearing cyclists while presenting a more condensed range of cyclist variants than the other two datasets, which have a broader pool of cyclists but a lower frequency of helmet-wearing cases. These datasets will be used in this study for training and testing detection models, and the campus testset will be useful for the evaluation of the tracks. A snapshot of the location and camera perspectives can be seen in Fig 3.

2.1.1 Camera setup

This research requires the creation of a multi-camera dataset with synchronized footage. The synchronization of the multi-camera dataset was achieved through a post-processing approach. Specifically, the first frame in which a visual clue became evident in each camera's footage was selected as the reference point for synchronization. This methodology ensured temporal alignment across the cameras, allowing for subsequent analysis and interpretation of the synchronized data. The cameras utilized are GoPro Hero 7 cameras, capturing 4k images every half second. The cameras will be set to take time-lapse images at every half second. To maintain consistency in the camera setup, all cameras were positioned to focus on a shared central point, maintaining a fixed distance of 4 meters from that center point. Specifically, two cameras were oriented towards each other, but their positions were adjusted to be situated 1 meter away from the side of the bike lane, while still maintaining a 4-meter distance from the central point. The third camera was positioned to maximize its perpendicular alignment relative to both cameras, also maintaining a 4-meter distance from the central point. This configuration (see Fig. 2) ensured a standardized and synchronized perspective for capturing cyclist and helmet data across the datasets.

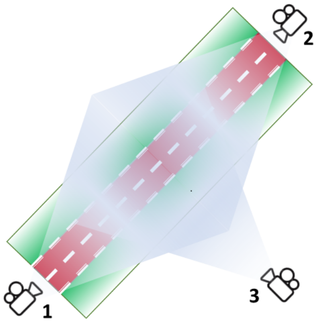


Fig. 2: This shows the general camera setup used to capture the data that we use in the datasets.

2.2 Object Detection

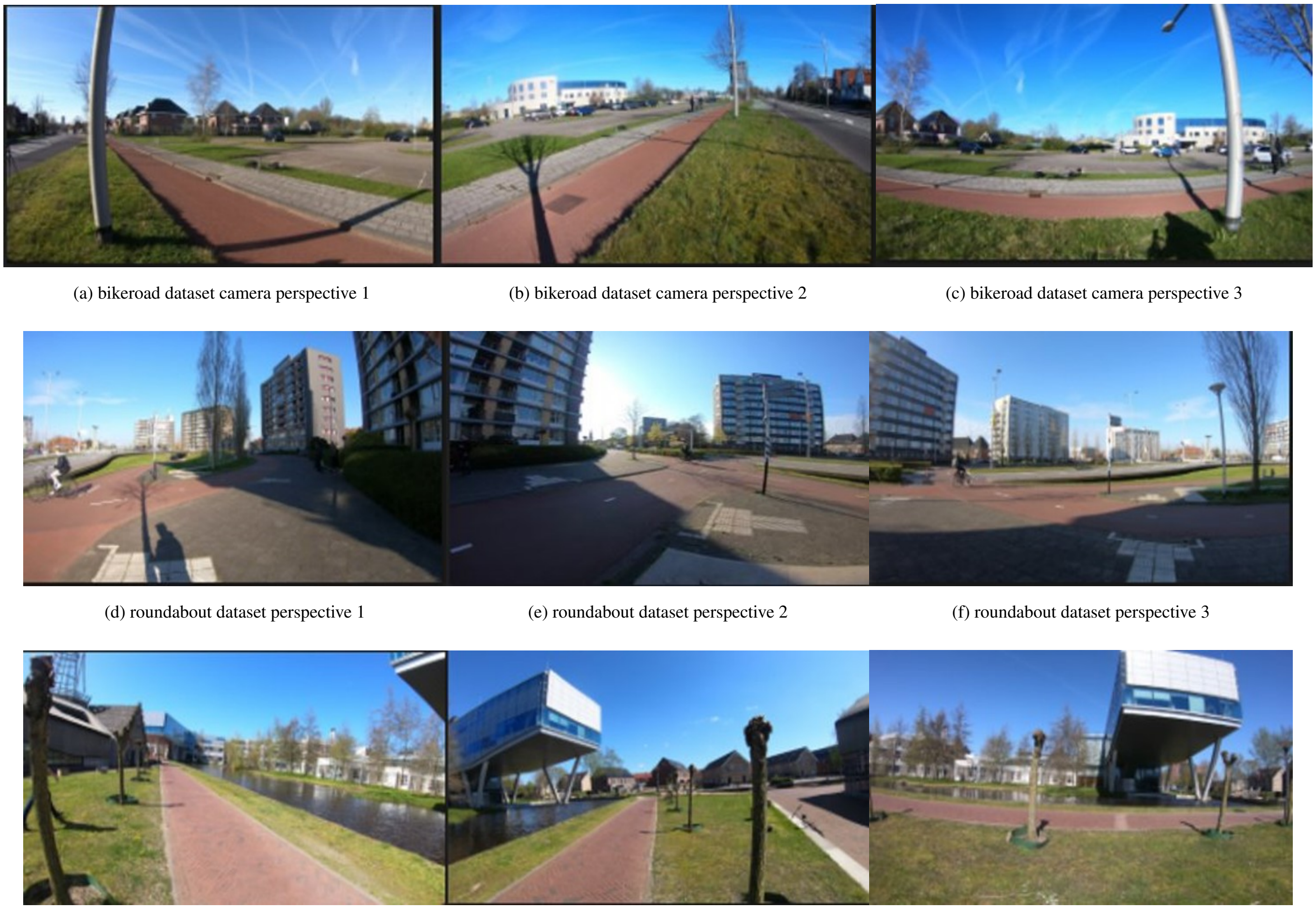
Yolov5 [21] is a state-of-the-art object detection algorithm that builds upon the previous success of the YOLO (You Only Look Once) family of models. It is a one-stage detector that achieves high accuracy and inference speed, making it a popular choice in real-time applications. Yolov5 uses a deep neural network to predict bounding boxes and class probabilities directly from an input image. It introduces several improvements over previous versions, including a new backbone architecture, a more efficient anchor box assignment strategy, and an improved loss function. These improvements result in higher accuracy and faster training times compared to previous YOLO models. The trained Yolov5 models will be used to detect the cyclists and helmets in our dataset and be evaluated in the experiments.

2.3 Tracking

The tracking approach utilized in this study is based on Norfair [22] and an application of SORT (Simple Online and Realtime Tracking) [17]. Norfair is composed of four components, including detection, estimation, data association, and track-ids. In each frame, a detection algorithm detects objects, and the estimation component predicts the location of the objects in the subsequent frame, which is also called an estimation. If the prediction matches a detection in the next frame the detections become a series called a track. This estimation is done by a Kalman filter module which uses a linear velocity module to represent the motion of each seperate object. The Kalman filter is initialized with zero velocity in any direction and with a very high uncertainty. The predict step predicts the next state of the track based on the previous bounding box and the update step estimates the system's current state based on the measurement at that time step.

To determine if there is a match, SORT uses a hungarian algorithm for its data association. Unlike SORT, which uses a hungarian algorithm for data association, Norfair utilizes a custom algorithm for matching predictions and estimations. This approach takes into consideration cases where minimizing the global minimum distance may not yield optimal results. The purpose of this approach is to prevent the algorithm from excessively prioritizing the minimization of global distance, which could lead to erroneous matches between objects that should not be associated. Instead, Norfair uses a matching strategy that involves sequentially selecting predictions and estimations based on their overlap, starting from the global minimum and progressing to the second minimum, and so forth. This sequential selection process continues until the minimum exceeds the maximum threshold, at which point we no other predictions and objects to be a match will be considered a match.

The last aspect of the tracker that needs to be discussed is the management of the tracks. For each successful detection match, a counter associated with a tracked object is increased. This counter increases by one when the association algorithm denotes 2 objects to be a match. On the other hand, we decrease the counter by one when a frame does not produce a match. The object is regarded as active



(a) bikeroad dataset camera perspective 1

(b) bikeroad dataset camera perspective 2

(c) bikeroad dataset camera perspective 3

(d) roundabout dataset perspective 1

(e) roundabout dataset perspective 2

(f) roundabout dataset perspective 3

(g) Campus dataset perspective 1

(h) Campus dataset perspective 2

(i) Campus dataset perspective 3

Fig. 3: The dataset for all locations showing the 3 different perspectives

once it has reached the initialization delay, which denotes a sufficient number of matches. With each succeeding match, the active object's counter adds hits to its total until it reaches the counter's predetermined upper limit. During tracking of an active object, if the counter reaches zero, the object returns to the RE-ID phase. The object will wait during this stage up to a number of frames to be matched with one of the initializing objects. If a match is found, the younger object is merged into the older one and then deleted from the tracking system. Otherwise, the older object is reactivated. Tracks that do not get matched for a certain amount of frames get removed. The states are visualised in Figure 4.

Additionally, We merged the helmet detections with the cyclist class during tracking. This was done to minimize the amount of tracking that is needed and due to the Kalman filter, which has difficulties with sudden and abrupt movements. Helmets relative to cyclists are quite small and could move in unsuspecting ways. This constraint was handled using a heuristic. A helmet was considered to be attached to a cyclist if its center point fell within the boundaries of the cyclists middle and highest points in the vertical plane as well as the cyclists width in the horizontal plane.

2.4 RE-ID

For the Re-identification we make use of the torchreid [25] research. This is an algorithm that extracts features from bounding boxes and use these features to make an estimation on the object seen from other camera setting. In this paper we will use torchreid to determine if objects obtained from different cameras are the same object. Models used to get the features are pretrained classifiers such as Resnet50, Densenet, inception and Xception and OSNet. In this study we will utilize OSNet to create the embedding and use the torchreid approach to compare the embeddings with each other. We will use OSNet since it has been specially trained for RE-ID purposes.

OSNet [26] is a lightweight model that has been specifically made for Re-identification using RE-ID networks. In order to obtain these multiscale features, OSNet employs various convolutional layers and then introduces an aggregation gate, which fuses the multiscale features with channel-wise weights to produce an embedding with the shape of 1 x 2048. We will use these computed embeddings in our research to compare them with one another. To do that we calculate the squared euclidean distance between the objects that could be match. By expanding the matrix', it is possible to compare two tracks of different lengths to one another. To ensure reliable matching, we impose a restriction where tracks must have an overlap in at least one frame, utilizing the synchronized input. For example, if tracks 6 and 24 are present in frame 70 but captured by different cameras, their embeddings are extracted, and the squared Euclidean distance between them is calculated. This process is repeated for all overlapping tracks captured by different cameras.

To establish one-to-one track matching between cameras, we employ the Hungarian algorithm. This algorithm minimizes a cost function to determine the optimal pairing of tracks between cameras. By minimizing the total embedding distances, each track in one camera is effectively matched to a track in the other camera. This approach enables us to establish correct associations between tracks from different camera perspectives, contributing to a comprehensive understanding of cyclist and helmet.

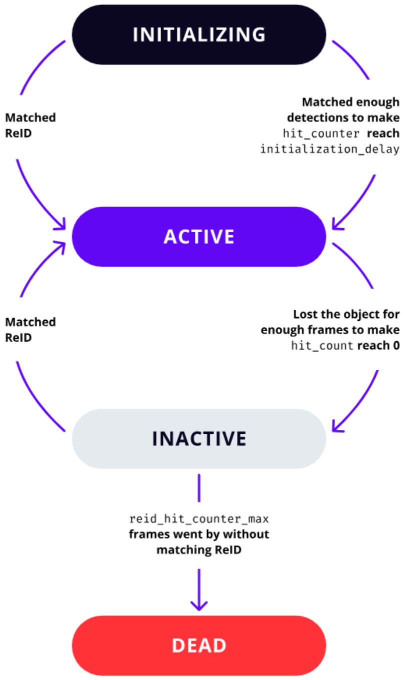

INITIALIZING
Matched ReID
Matched enough detections to make hit_counter reach initialization_delay
ACTIVE
Matched ReID
Lost the object for enough frames to make hit_count reach 0
INACTIVE
reid_hit_counter_max frames went by without matching ReID
DEAD

Fig. 4: The process an object goes through once it gets detected and initialized. [24]

2.5 Metrics

In computer vision, evaluating the performance of object detection models requires the use of several metrics. In this paper, we employ precision, recall, and the F1 score to quantify the effectiveness of our object detection model.

Precision, also known as positive predictive value, measures the proportion of correctly (with an overlap with the ground truth of 50%) predicted positive instances out of the total instances predicted as positive . It is computed using the following equation:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{1}$$

Recall, also known as sensitivity or true positive rate, assesses the proportion of correctly predicted positive instances out of the total actual positive instances. The equation for recall is as follows:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

The F1 score is a harmonic mean of precision and recall, providing a single metric that balances both measures. It is calculated using the following equation:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

These metrics collectively provide a comprehensive evaluation of the object detection performance, taking into account both the precision and recall aspects. The precision metric emphasizes the accuracy of positive predictions, recall highlights the ability to detect positive instances, and the F1 score combines both measures to offer a balanced assessment.

The evaluation of tracking results involves the classification of potential errors between predicted and ground-truth tracks into three distinct types: detection errors, localization errors, and association errors. [27].

Detection errors refer to situations where the tracking algorithm fails to detect an object or incorrectly identifies a non-existent object. These errors can occur when the object is occluded, partially visible, or has a low contrast with the background.

Localization errors occur when the predicted bounding box or position of the object deviates significantly from the ground truth. Such errors can arise due to inaccurate estimation of object boundaries, perspective distortions, or imprecise localization algorithms.

Association errors occur when the tracking algorithm incorrectly assigns detections to tracks or fails to associate detections with the correct tracks. These errors can occur when objects are in close proximity, undergo significant appearance changes, or exhibit similar motion patterns.

In order to calculate these errors, we employ a set of established tracking metrics. Firstly, the Multi Object Tracking Accuracy (MOTA) and Multi Object Tracking Precision (MOTP) metrics are calculated. These metrics are widely used for assessing the performance of multiple object tracking algorithms and are described in the clearMOT metrics framework [28]. Additionally, we evaluate the tracker's output based on the IDentity f1 (IDF1) score, as defined in [29]. First we will look into how MOTA and MOTP are calculated and what tracking error they evaluate. after that we will look at how IDF1 does it differently and

MOTA primarily focuses on detection performance, considering a bijective one-to-one mapping between the ground truth set (GT) and the predictions. True Positives (TP) are the predictions that correctly match the ground truth, while False Positives (FP) correspond to predictions without any matching ground truth, and False Negatives (FN) represent ground truth instances without corresponding predictions. The tracking accuracy is then computed by accounting for these ID switches (IDS). They occur when the tracker mistakenly assigns the different ID to a ground truth track then previously was assigned to it. The MOTP metric complements MOTA by specifically considering localization errors, providing a comprehensive evaluation of the tracker's performance using the MOT metrics [28].

MOTA quantifies the tracking accuracy at the detection level, with a higher value indicating better performance. It assesses the percentage of false positives, false negatives, and mismatches in the tracking results of a multiple object tracking algorithm. See Equation 4

$$MOTA = 1 - \sum_i \frac{FN_i + FP_i + IDS_i}{GT_i} \tag{4}$$

where FN_i is the number of false negatives in image i, FP_i is the number of false positives, IDS_i is the number of identity switches or mismatches in image i, and GT_i is the number of ground truth objects in image i.

MOTP measures the average similarity error S between the predicted and ground truth object locations, indicating the tracking precision of the algorithm. A higher value indicates better tracking precision. See Equation 5

$$MOTP = \frac{1}{TP_t} \sum_i S_i \tag{5}$$

where TP_t is the number of True Positives over all the images $_t$, S is the similarity score for objects in image i. In our research the score is calculated as a intersect over union (IoU). The output results in the average IoU over all correct predictions.

In order to calculate the IDF1 score, the ground truth trajectories and predicted trajectories are matched in a one-to-one manner. This matching process is performed using the Hungarian algorithm, which calculates the number of true positives (TP), false positives (FP), and false negatives (FN). The goal is to minimize the occurrence of FP and FN resulting from the trajectory matching.

Once the matching is completed, the metrics of identity true positive (IDTP), identity false positive (IDFP), and identity false negatives (IDFN) are determined. These metrics are then used to calculate the identity recall (IDR), identity precision (IDP), and identity F1 score (IDF1). The IDR represents the proportion of correctly identified trajectories out of all the ground truth trajectories. The IDP indicates the accuracy of the predicted trajectories in terms of correctly identifying the ground truth. The IDF1 score combines both recall and precision, providing a single measure to evaluate the overall performance of the trajectory matching process [29]. The Equations used to get to these numbers are here Equations 6, 7, 8.

$$IDR = \frac{IDTP}{IDTP + IDFN} \tag{6}$$

$$IDP = \frac{IDTP}{IDTP + IDFP} \tag{7}$$

$$IDF1 = \frac{IDTP}{IDTP + 0.5 * IDFP + 0.5 * IDFN} \tag{8}$$

3 Experiments

The objective of the first experiment is to investigate the impact of utilizing different cameras for detecting cyclists and identifying helmets. For this experiment a Yolov5s model is trained on each camera for each location separately. This way we can compare the performance of the cameras to each other. We will evaluate the results on the campus testset consisting of multiple cyclists, some wearing helmets and others not. For the evaluation we employed the precision, recall and F1-score metrics described in the Metrics section.

To assess the effectiveness of the models tracking performance, a second experimental setup has been devised. The primary aim is to accurately track all passing cyclists, as well as associating each helmet with the corresponding cyclist. This tracking process will utilize the individual video frames and the bounding boxes obtained by the trained object detection models. Additionally we generate tracking bounding boxes around the detected cyclists in the images. This way we can qualitatively evaluate the performance of the model as well. The evaluation of detected helmets will be done in the later experiment. To assess the effectiveness of the models and evaluate tracking performance quantitatively. The resulting tracks are then evaluated individually, employing the dedicated evaluation metrics as described in the materials and methods section.

The third experiment is dedicated to Re-identification (RE-ID), with the objective of linking tracks generated from the tracker with different cameras perspectives in a one-to-one manner. The main aim is to establish corresponding track numbers for successful linking. Through this linking process, the number of detected helmets within the linked tracks can be extracted. The video output of the tracker plays a critical role in this step, enabling the manual matching of track IDs from different cameras to establish the most appropriate links between tracks. This manual linking process serves as the ground truth for the experiment. The results of RE-ID will undergo thorough analysis and evaluation to assess the overall effectiveness of the framework, including detection, tracking, and RE-ID. Specifically, the analysis will focus on correctly detecting the number of passing cyclists and determining whether the linking of tracks has improved the prediction of cyclists wearing helmets also a prediction of helmet will be made based on the amount of frames the track are detected during this experiment we also look in increments of 10% which threshold yields the best performance.

3.1 Results

This subsection presents a comprehensive overview of the experiments carried out, outlining the applied methodology as described in the previous sections. The results obtained from the experiments are briefly summarized, highlighting the capabilities of object detection, tracking, and Re-identification (RE-ID). Notably, special attention is given to the RE-ID process, as it directly addresses the second research question.

3.1.1 Object detection

The results of the object detection experiments are presented separately for each location: campus, roundabout, and bike-road (as shown in Table 2). The initial hypothesis presumed that the similarity in camera positions and road lanes across different locations, as depicted in Figure 2, would result in comparable performance of the trained models in detecting cyclists and helmets across the various locations. Nevertheless, the experimental results contradict this assumption.

Table 2: Results of object detection on different datasets

Yolov5s	Campus Dataset		
Camera	Precision	Recall	F1 score
cam1	45.4%	6.3%	11.1%
cam2	71.3%	55.3%	62.3%
cam3	97.9%	77.3%	86.3%
	Roundabout Dataset		
Camera	Precision	Recall	F1 score
cam1	16.1%	7.8%	10.5%
cam2	93.9%	18.5%	30.9%
cam3	96.2%	38.4%	54.9%
	Bikeroad Dataset		
Camera	Precision	Recall	F1 score
cam1	97.9%	3.3%	6.3%
cam2	90.7%	17.8%	29.8%
cam3	98.4%	26.0%	41.1%

In the presented table, it is evident that the model for camera 3 exhibits the highest performance in terms of precision, recall, and F1-score. Specifically, the campus dataset demonstrates the highest F1-score, suggesting that the model performs most effectively when detecting objects from a perpendicular view. Camera 2, on the other hand, shows a slightly lower performance compared to camera 3 in the roundabout dataset, with a 3% difference. However, camera 2 exhibits a considerably low recall of 18.5%, resulting in a F1-score of 30.9%. This pattern is also observed in the bikeroad dataset, where the precision stands at 90.7% but the recall is notably low at 17.8%. These findings indicate that while the model's predictions are mostly accurate, it fails to detect a significant number of objects that are indeed present

In the results section, it was evident that camera 1 displayed the lowest F1-score across all datasets, accompanied by consistently low recall values, for the campus roundabout and bikeroad, 6.3%, 7.8% and 3.3% respectively. Consequently, camera 1 was deemed unreliable for further analysis due to its unexpected and inconsistent performance. The subsequent RE-ID experiments focused exclusively on cameras 2 and 3. The decision to exclude camera 1 was prompted by the observed anomaly, which could potentially be attributed to a rotational offset of approximately 30 degrees clockwise in the testing dataset.

It is worth noting that in addition to the object detection findings, a noteworthy observation is the slightly higher recall achieved in the roundabout dataset compared to the bikeroad dataset. This observation is likely attributed to the greater variation in angles, distances, and lighting conditions present in the roundabout dataset.

3.1.2 Tracking Results

In the 2nd experiment the aim is to track the cyclist objects that the object detector detects. Therefore we employ the tracker described in the methology on the campus dataset. These results will be evaluated using the tracking evaluation metrics described in the Metrics section. The resulting Metrics by the tracker can be seen in Table 3.

In the table you can see that the IDF1 score for the tracking in the frames captured by camera 2 is 34.5% and for the frames captured with camera 3 is 56.4%. This shows that the tracks in camera 3 can be more easily matched with the ground truth. The IDP of camera 3 is 80.2% which indicates that the predicted bounding boxes on average

Table 3: Results of the tracking on the campus dataset.

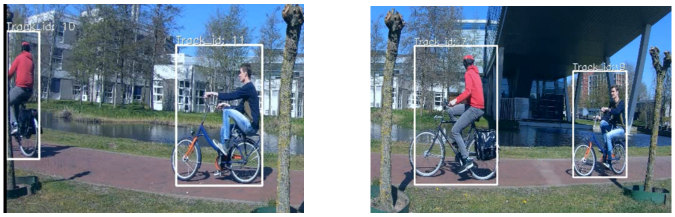
Campus	IDF1	IDR	IDP	MOTA	MOTP	precision	recall	Idsw	Fragmented
cam 1	1,6%	1,1%	100,0%	0,9%	49,0%	100,0%	1,1%	1	1
cam 2	34,5%	34,3%	34,7%	18,5%	67,9%	60,0%	59,2%	39	50
cam 3	56,4%	45,1%	75,7%	54,9%	80,2%	96,4%	57,3%	5	4

match for 80.2% with the ground truth. Comparing this to camera 2 with 67.0% shows that the model in camera 3 can more precisely detect cyclists in the frames. Additionally the MOTA for the model for camera 3 with a performance of 54.9% shows an increase compared to the model for camera 2, which has a MOTA of 18.5%

An analysis of the generated video outputs revealed that camera 2 encountered numerous IDS and fragmentations compared to camera 3, 39 IDSW and 50 fragmentation compared to 5 IDS and 4 fragmentations. These issues arose primarily in the distant background, as a result of pixelated object representations and frequent occlusions. Consequently, the tracks often disappeared and reappeared in later frames, preventing successful connection through the association component within the tracker. A specific example illustrating the occurrence of identity switches in camera 3 can be seen in Fig. 5. Furthermore, an example showcasing both the fragmentations and identity switches in camera 2 can be observed in Fig. 6.

These findings underscore the challenges associated with the performance of camera 2 in the tracking process, primarily due to the specific characteristics of its background and the resulting occlusions. On the other hand, camera 3, demonstrated more stable tracking performance and exhibited fewer issues related to identity switches and fragmentations.

Overall, the results from the tracking experiments validate the previous observations made during the object detection phase and provide insights into the specific challenges and limitations encountered by each camera during the tracking process.



(a) 2 Detected cyclists with respectively track id 10 on the left and 11 on the right

(b) 2 Detected cyclists with respectively track id 7 on the left and 9 on the right

Fig. 5: The above images show the IDS happening in camera perspective 3. The cyclists move over the road from right to left. The tracker losses the objects behind the tree for a moment and assigns a new id to the same object.

3.1.3 Re-identification

The results of the RE-ID experiment are presented in Table **??**. This Table shows the linked tracks connected by the Hungarian algorithm based on the embedding distanced computed and extracted by the RE-ID software. The table also shows the prediction whether the cyclist wears a helmet. The threshold of 40% is shown, other records can be found in Appendix 6. According to Table **??**, a total of 10 cyclists were detected passing by, with 5 of them wearing helmets. However, the ground truth is that there are 9 cyclists passing by, of which 6 were wearing helmets. The presence of additional tracks can be attributed to identity switches, whereby a single ground truth track is associated with multiple detection track IDs. Regarding the erroneous prediction regarding the helmet status occurred in track 8, the framework determines that the cyclist is not wearing a helmet, whereas in reality, the cyclist does wear a helmet. Table 5 provides quantitative results from the RE-ID experiment, offering further insights. This Table shows that in track-ID 1 is matched to track-id 14 where track-ID 1 appeared in 87 frames and got 58 helmet detections in those frames. For track-ID 14 the number of frames detected was 87 as well with only 37 helmet detection. This shows that the percentage helmets detected based on the total number of frames is lower in camera 2 for the linked objects 1 and 14. This trend continues throughout all the links made by the RE-ID software. Additionally it is good to note that when a cyclists doesn't not wear a helmet in both linked track-ids the number of detected helmets is also low. as can be seen in both track-ID 11 and 12. These tracks link with track-IDs 44 and 49 which have all helmet detections below 3.

4 Discussion

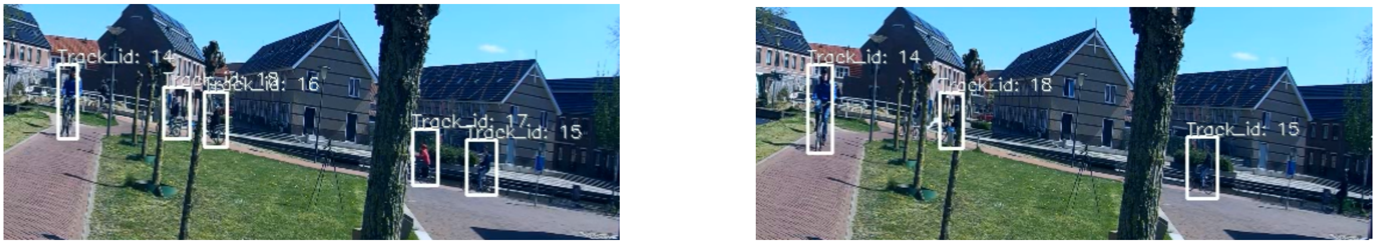
In this study, we researched a comprehensive framework utilizing YOLOv5 as a detector, along with Norfair and torchreid as tracking and Re-identification systems, respectively. This framework is utilized to detect, track, and Re-identify cyclists and helmets across multiple cameras. Our results demonstrated the effectiveness of this framework in counting cyclists and determining the percentage of helmet usage. During the course of our research, an important finding emerged concerning the performance of the model assigned to camera 2. It was observed that this particular model yielded less favourable results than anticipated, this was likely due to the significant occlusions encountered throughout the detection and tracking process. This finding highlights the impact of occlusions on the tracking performance of the model associated with camera 2. The findings from the Re-identification process indicate that accurate helmet detection percentages were achieved through successful linking of the correct tracks. The analysis highlights the substantial contribution of the model associated with camera 3 in this process. Furthermore, the model assigned to camera 2 played a validating role, supporting the results obtained from the camera 3 model. The consistent trend observed in the findings reveals a lower count of helmets when cyclists are not wearing them, and conversely, a higher count when helmets are present. This consistency reinforces the effectiveness of the framework in accurately identifying the presence or absence of helmets.

These findings demonstrate the potential of the proposed framework for cyclist detection and identification tasks. Moving forward, it is crucial to focus on improving the tracking capabilities of the framework. This could involve enhancements in the diversity an availability of cyclist and helmet datasets or to the tracking RE-ID algorithm to improve the tracking through occlusions or the implementation of measures to minimize the field of view and the potential for occlusions. Furthermore, this work sets the stage for future advancements aimed at refining tracking capabilities and expanding datasets to ensure effective cyclist and helmet detection in different locations.

5 Conclusions

In this study, our aim was to develop a framework for detecting cyclists and identifying their helmet usage using object detection, tracking, and Re-identification techniques. We employed a Yolov5 model to detect cyclists and helmets from three different camera perspectives, with the perpendicular view yielding the most favorable results. However, challenges were encountered with the second camera perspective, primarily attributed to occlusions and distant objects leading to pixelation.

In the final experiment, we implemented Re-identification between the second and third camera perspectives. The results demonstrated that after association, the model exhibited a small deviation of only



(a) Multiple detected object objects in the 2nd camera perspective

(b) Multiple detected object objects in the 2nd camera perspective

Fig. 6: The images above show an example of the id switching. All these cyclists go right to left. An id switch can be seen in image 6b. Here, object 18 is reconnected with the cyclist with trackid 16 in the earlier image 6a. Another observation you can make is that the cyclists with track id 17 and 16 in image 6a are currently occluded and result in fragmentations since they wont reconnect to any other objects and the ground truth track havent finished

Table 4: The Table below shows track ids detected in both cameras. The track ids in the same column are matched by the Hungarian algorithm to be the same cyclist. The last row signifies whether the cyclist has a helmet.

Campus	**Associated track_ID**									
cam3	1	2	4	5	7	8	10	11	12	13
cam2	14	15	22	29	28	21	30	44	49	51
Helmet prediction 40%	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
Correct Match	Correct	Correct	Correct	Correct	Correct	Correct	Miss	Correct	Correct	Correct
Correct Helmet prediction	Correct	Correct	Correct	Correct	Correct	False	Correct	Correct	Correct	Correct

one incorrect track, successfully generating nine correct tracks out of a total of ten. Additionally, the framework accurately identified the number of helmets in nine out of the ten tracks, further highlighting its efficacy.

Overall, our study presents a comprehensive framework for cyclist detection, tracking, and Re-identification. While certain limitations were identified, such as occlusions impacting tracking performance, our proposed framework showcases promising accuracy in detecting cyclists and determining their helmet usage. Further refinements and enhancements to address the identified limitations could lead to improved performance and practical applications in cyclist safety and monitoring systems.

6 Future work

To further enhance the performance and applicability of the framework, several areas for future work can be identified. First, expanding the dataset to include diverse environmental conditions, cyclist behaviors, and camera perspectives can improve the generalization capabilities of the object detection models. This dataset expansion will ensure the robustness of the framework across various real-world scenarios. Efforts should also be directed towards developing advanced techniques to handle occlusions during the tracking process. Exploring multi-object tracking algorithms capable of robustly dealing with occlusions can significantly improve tracking accuracy and reduce fragmentation and id switches.

Refining the Re-identification algorithms is crucial for addressing challenges associated with occlusions and fragmented tracks. Including spacial information, cyclist trajectory and appearance consistency, can enhance the accuracy of Re-identification and reduce false associations. Moreover, optimizing the framework for real-time implementation in live monitoring systems or intelligent traffic management is essential. Exploring hardware acceleration options can facilitate efficient real-time processing and enable the framework to perform effectively in real-world scenarios.

Finally, investigating the integration of the proposed framework with existing cyclist safety systems can provide insights into its potential impact on cyclist safety and accident reduction.

Acknowledgements

- This project is a collaboration between the Computer Vision & Data Science professorship and Provinsje Fryslân within the ROF (Regionaal Orgaan verkeersveiligheid Fryslân) programme.

provinsje fryslân
provincie fryslân

References

[1] Maarten Kroesen. To what extent do e-bikes substitute travel by other modes? evidence from the netherlands. *Transportation Research Part D: Transport and Environment*, 53:377–387, 2017.

[2] fietsdoden, 2021. https://www.cbs.nl/nl-nl/maatwerk/2022/37/fietsdoden-2021. accessed: 16-03-2023.

[3] Meer fietsdoden na eenzijdige ongevallen. https://www.cbs.nl/nl-nl/nieuws/2022/37/meer-fietsdoden-na-eenzijdige-ongevallen. accesed: 16-03-2023.

[4] Crispijn L. van den Brand, Lennard B. Karger, Susanne T.M. Nijman, Huib Valkenberg, and Korné Jellema. Bicycle helmets and bicycle-related traumatic brain injury in the netherlands. *Neurotrauma Reports*, 1(1):201–206, 2020.

[5] Fietshelm — fryslan. https://www.fryslan.frl/fietshelm. accesed: 16-03-2023.

[6] Romuere Silva, Kelson Aires, Thiago Santos, Kalyf Abdala, Rodrigo Veras, and André Soares. Automatic detection of motorcyclists without helmet. In *2013 XXXIX Latin American computing conference (CLEI)*, pages 1–7. IEEE, 2013.

[7] Aihua Zheng, Lei Zhang, Wei Zhang, Chenglong Li, Jin Tang, and Bin Luo. Local-to-global background modeling for moving object detection from non-static cameras. *Multimedia Tools and Applications*, 76(8):11003–11019, Apr 2017.

[8] Bao Zhimin Wang Xiao Tang Jin Li, Chenglong. Moving object detection via robust background modeling with recurring patterns voting. 2018.

[9] Romuere R. V. e. Silva, Kelson R. T. Aires, and Rodrigo de M. S. Veras. Detection of helmets on motorcyclists. *Multimedia Tools and Applications*, 77(5):5659–5683, Mar 2018.

[10] C. Vishnu, Dinesh Singh, C. Krishna Mohan, and Sobhan Babu. Detection of motorcyclists without helmet in videos using convolutional

Table 5: The Table below shows the associated track ID and their respective counts in the number of frames the track was detected and how often a helmet was detected. The prediction made in Table **??** is determined if the combined number of helmets exceeds 40% of the combined numbers of frames.

Track-ID	**1**	**2**	**4**	**5**	**7**	**8**	**10**	**11**	**12**	**13**
Helmets	58	120	158	97	0	40	1	2	0	75
Frames	87	166	226	220	72	57	61	100	94	101
Track-ID	**14**	**15**	**22**	**29**	**28**	**21**	**30**	**44**	**49**	**51**
Helmets	37	85	0	11	49	92	0	1	0	70
Frames	87	259	76	21	198	343	18	146	10	135

neural network. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3036–3041, May 2017.

[11] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021.

[12] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. *CoRR*, abs/1702.04593, 2017.

[13] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, Feb 2008.

[14] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. *CoRR*, abs/2007.07247, 2020.

[15] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4256–4265, June 2016.

[16] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification, 2019.

[17] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2016.

[18] A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[19] Harold W. Kuhn. *The Hungarian Method for the Assignment Problem*, pages 29–47. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[20] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.

[21] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios, 2021.

[22] github - tryolabs/norfair: Lightweight python library. https://github.com/tryolabs/norfair. accesed: 16-03-2023.

[23] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council.

[24] Javier Berneche. Norfair 2.0, 2022.

[25] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch, 2019.

[26] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification, 2021.

[27] Ido Leichter and Eyal Krupka. Monotonicity and error type differentiability in performance measures for target detection and tracking in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2553–2560, 2013.

[28] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, Feb 2009.

[29] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking, 2016.

Table 6: Calculation to determine which threshold yields the best results

Calculation to determine which threshold yields the best results											
cam3	1	2	4	5	7	8	10	11	12	13	
cam2	14	15	22	29	28	21	30	44	49	51	**number of cylcists with a helmet**
10%	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	**7**
20%	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	**6**
30%	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	**6**
40%	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	**5**
50%	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	**3**
60%	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	**1**
70%	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**0**
80%	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**0**
90%	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**0**
100%	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**0**

A helmet Detection