Comprehensive Analysis of YOLOv8 + DeepSORT for Vehicle Tracking: HOTA and CLEAR-Based Evaluation

I Nyoman Eddy Indrayana[1], Made Sudarma[2], I Ketut Gede Darma Putra[3], Anak Agung Kompiang Oka Sudana[4]
Engineering Science Department, Udayana University, Badung, Indonesia[1]
Electrical Engineering Department, Udayana University, Badung, Indonesia [2]
Information Technology Department, Udayana University, Badung, Indonesia[3, 4]

***Abstract*—This paper offers a thorough comparative investigation of the performance of a vehicle multi-object tracking system, incorporating various versions of the YOLOv8 detector (from 'n' to 'x') alongside the DeepSORT tracking algorithm. This study systematically assesses the impact of the trade-off between detector speed and accuracy on tracking metrics, utilising a real-world traffic video dataset from Bali. The assessment is performed utilising two fundamentally distinct metric frameworks: the traditional CLEAR metric (which includes MOTA) and the contemporary Higher Order Tracking Accuracy (HOTA) metric. The findings indicate that although the larger YOLOv8 model markedly enhances detection recall, particularly for smaller and more difficult items like motorcycles, tracking issues persist. The dual metric study provides significant insights: the HOTA measure demonstrates that car tracking has more associative stability (higher AssA scores) compared to motorbike tracking, which frequently experiences track fragmentation. In contrast, the detection-biased MOTA metric produces somewhat paradoxical outcomes, as motorbikes receive elevated scores due to enhanced detection accuracy (fewer false positives), therefore obscuring deficiencies in tracking consistency. This study concludes that HOTA offers a more comprehensive evaluation by differentiating between detection and association performance, so demonstrating that detection-only metrics like MOTA can yield an imperfect representation of actual tracking ability. These findings underscore the necessity of matching detector architecture and evaluation criteria with specific application requirements, particularly in safety-critical systems where identity consistency is essential.**

Keywords—Multi-object tracking; higher order tracking accuracy metric; CLEAR Metric; YOLOv8

I. Introduction

Intelligent Transportation Systems (ITS) have emerged as a fundamental component in worldwide initiatives aimed at enhancing the efficiency, safety, and sustainability of transportation infrastructure [1]. Central to these systems is the capacity to dynamically observe, analyse, and regulate vehicle movements [1]. Multiple Object Tracking (MOT) has become a pivotal technology facilitating these sophisticated applications [1], [2]. The applications of MOT in Intelligent Transportation Systems are varied and significant, encompassing traffic flow analysis for congestion management, automated traffic enforcement, data collection for urban planning, and accident detection [3], [4]. The cornerstone of all these sophisticated systems is the accessibility of precise, real-time data regarding the movement of each vehicle on the road [1], [3].

The predominant and effective methodology in contemporary Multi-Object Tracking (MOT) is the tracking-by-detection paradigm [5]. This paradigm deconstructs the intricate tracking issue into two more feasible steps: initially, identifying all objects of interest (namely, cars) in each video frame; and subsequently, linking (associating) these detections across time to establish a coherent trajectory for each object [6]. The efficacy of this method is largely contingent upon the functionality of these two elements [7]. The integration of the YOLO (You Only Look Once) detector family with the SORT (Simple Online and Realtime Tracking) tracker family has gained significant popularity [8], showcasing the efficacy and efficiency of this paradigm across diverse real-world applications [8], [9].

Although several investigations have utilised the integration of YOLO and DeepSORT, most choose to employ a singular version of YOLO (e.g. YOLOv5) [10] and concentrate on enhancing the tracking algorithm itself. A comprehensive comparison review of the whole range of YOLOv8 devices, from the lightweight nano form to the robust extra-large variant, is still absent, particularly with on-road vehicle tracking.

Significantly, numerous current evaluations continue to depend extensively on traditional criteria like MOTA (Multiple Object Tracking Accuracy) [11], [12], [13], which are recognised for their substantial bias against detection performance. These measurements frequently conceal association failures (identification tracking) by aggregating them with detection mistakes [6]. The influence of detector scale selection on contemporary metrics like HOTA (Higher Order Tracking Accuracy), which distinctly differentiate between detection and association assessment, has not been extensively investigated [6].

This prompts a fundamental research inquiry: How does the compromise between detector precision and velocity (e.g. YOLOv8n versus YOLOv8x) affect overall tracking efficacy when assessed by a metric that equally prioritises detection accuracy and association consistency? The link between detection quality and tracking quality is hypothesised to be nonlinear. There may come a juncture where augmenting detector strength ceases to yield advantages or becomes counterproductive, since more sensitive detectors can generate partial or ambiguous detections that obfuscate the tracker's association logic [6], [14].

This evaluation aims to identify specific trends in vehicle behaviour that may lead to accidents. Consequently, the findings of this research will enhance vehicle monitoring technology and facilitate the design of more efficient traffic safety systems. This research seeks to elucidate the correlation between vehicle trajectories and accident detection, as well as to formulate algorithms that are more responsive to fluctuating road circumstances.

This study significantly contributes to the fields of computer vision and intelligent transportation systems through the following aspects:

- Comprehensive Empirical Evaluation: Provides an in-depth empirical evaluation of five YOLOv8–DeepSORT combinations on a challenging and contextually pertinent real-world vehicle monitoring dataset gathered from Bali, Indonesia.
- Dual Metric Analysis: Performs a thorough performance evaluation utilising both the CLEAR metric and the more extensive HOTA metric, offering a detailed perspective on tracker behaviour that is unattainable by a singular metric alone.

The subsequent sections of this work are structured as follows. Section II comprises a literature review that elucidates pertinent studies, emphasizing advancements in object identification and tracking, specifically with YOLO and DeepSORT. Section III delineates the proposed approach and the materials utilized, encompassing dataset preparation, model configuration, and evaluation protocol. Section IV presents the experimental data and offers a comparative analysis utilizing the HOTA and CLEAR metrics. Section V presents the conclusions of this article and delineates prospective avenues for further research.

II. Literature Review

A. Advancement of Object Detection Utilising YOLO

Object detection is a crucial component in numerous computer vision applications, encompassing object tracking. The YOLO (You Only Look Once) architecture, introduced in 2016 with YOLOv1, represented a substantial advancement in real-time object recognition with an efficient one-stage detection methodology [15]. The progression of YOLO persists with enhancements in design and training methodologies in YOLOv2 [16], YOLOv3 [17], YOLOv4 [18], YOLOv5 [19], YOLOv6 [20], and YOLOv7 [21]. Ultimately, YOLOv8 provides a flexible framework featuring an extensive array of models tailored for diverse computational and accuracy needs [22], [23], [24].

The progression of YOLO (You Only Look Once) from version 1 to 8 signifies a swift enhancement in real-time object identification. YOLOv1 initiated a paradigm shift by conceptualising detection as a singular regression task, directly forecasting bounding boxes and classes from the entire image in one iteration, hence achieving much greater speed compared to earlier two-stage detectors. Subsequent iterations, including YOLOv2 through YOLOv5, progressively enhanced this architecture by incorporating anchor boxes, more resilient backbones like CSPDarknet53 [25], and feature aggregation networks such as PANet to optimise the balance between accuracy and speed [26]. YOLOv8 achieved notable advancements by enhancing the backbone and neck for improved feature extraction and implementing a pivotal modification through the adoption of an anchor-free detection head [25]. This anchor-free methodology streamlines the prediction process and enhances the model's flexibility to objects of diverse scales and aspect ratios, distinguishing it from its predecessors that depended on specified anchor boxes.

Recent advancements indicate that augmenting YOLOv8 with supplementary modules can enhance performance in intricate circumstances. Abdullah N. Alhawsawi et al. introduced an Enhanced YOLOv8 using a Context Enrichment Module (CEM) to enhance crowd counting in drone imagery by more effectively differentiating small, dense targets[27]. This method enhances detection precision, while our research combines YOLOv8 with DeepSORT to maintain identity coherence in vehicle tracking. These studies demonstrate two alternative approaches: enhancing detector capability or integrating detection with robust tracking for dependable multi-object surveillance in intelligent transportation systems.

B. Intersection over Union (IoU)

In contemporary object detection systems, a crucial element that quantifies spatial accuracy between predictions and ground truth annotations is Intersection over Union (IoU). Intersection over Union (IoU) is a crucial assessment metric in numerous deep learning-based object detection algorithms [28], notably the YOLO (You Only Look Once) model family and its most recent iteration, YOLOv8. In the realm of vehicle recognition within roadway settings, pertinent to Intelligent Transportation Systems (ITS) applications or traffic surveillance, elevated spatial precision is essential to enable the system to accurately identify and detect cars across diverse environmental circumstances.

In the evaluation of object detection performance, Intersection over Union (IoU) is commonly employed as a metric to quantify the overlap between predicted and actual bounding boxes. The Intersection over Union (IoU) is determined by the ratio of the intersection area of the predicted and ground truth boxes to their total area. A high IoU value signifies a more precise prediction in object localisation. The Intersection over Union (IoU) is frequently employed as a criterion to ascertain whether a detection is classified as positive or negative in the computation of accuracy and recall metrics. The accuracy and precision of object detection, often assessed using IoU-based metrics, directly influence the efficacy of subsequent tracking tasks. Inaccurate detection may result in track fragmentation and misidentification during the tracking process. Intersection over Union (IoU) is the ratio of the overlapping area of two bounding boxes, specifically the model prediction and the ground truth annotation, to the total area of their union. The mathematical formulation of IoU is as follows:

$$IoU = \frac{|A_p \cap A_{gt}|}{|A_p \cup A_{gt}|} \quad (1)$$

The bounding box of the model prediction results is denoted as A_p, while the bounding box of the ground truth data is represented as A_{gt}, The symbol ∩ signifies the overlapping area (intersection), and ∪ denotes the total combined area of the two boxes (union).

In computational implementation, if each bounding box is denoted by its border coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$, the intersection area is computed as (2):

$$Area_{inter} = max\left(0, min\left(x_p^{max}, x_{gt}^{max}\right) - max\left(x_p^{max} - x_{gt}^{max}\right)\right) x\ max\left(0, min\left(y_p^{max}, y_{gt}^{max}\right) - max\left(y_p^{max} - y_{gt}^{max}\right)\right) \quad (2)$$

and the union area is computed as (3):

$$Area_{union} = Area_p + Area_{gt} - Area_{inter} \quad (3)$$

The IoU value spans from 0 to 1, with 1 signifying complete overlap and 0 denoting no overlap at all.

C. Object Tracking with DeepSORT

The SORT (Simple Online and Realtime Tracking) algorithm, developed by Bewley in 2016, provides an efficient solution for the multiple object tracking (MOT) task by employing object detection and a Kalman filter for movement prediction, alongside the Hungarian algorithm for associating detections across successive frames [8]. Nevertheless, SORT is significantly reliant on detection quality and is less proficient in managing variations in object appearance and occlusion. DeepSORT addresses this restriction by including appearance cues derived from a deep neural network designed for re-identification tasks. The utilisation of these visual attributes enables DeepSORT to enhance object tracking, even amongst alterations in appearance or transient occlusion [29].

DeepSORT is constructed upon the SORT framework, which utilises bounding box location estimation via a Kalman filter and employs the Hungarian method for data association [10], [11]. The primary innovation of DeepSORT is the utilisation of appearance information derived from objects through a Convolutional Neural Network (CNN) [30]. These visual attributes offer supplementary information during the association process, especially in preventing misidentification when two items overlap or momentarily vanish from the frame.

The tracking procedure utilising DeepSORT comprises three primary components. Initially, object identification is executed utilising detection architectures like YOLOv8, SSD, or Faster R-CNN. Subsequently, the item's position in the subsequent frame is forecasted employing a Kalman Filter. Third, an association process is conducted between the new detection and the existing track, using spatial metrics and visual similarities. This enables DeepSORT to preserve object identity with greater precision and demonstrates resilience to visual noise.

The Kalman Filter is essential for forecasting the position of the monitored item between successive video frames. This method is executed iteratively. This filter determines the object's position through the coordinates of the bounding box's centre (u, v), the aspect ratio (width divided by height is γ), the height of the bounding box (h) and the predicted rate of change for each of these components $(\dot{u}, \dot{v}, \dot{\gamma}, \dot{h})$. Utilising this predictive capability, DeepSORT can adeptly forecast the object's movement within the video. The state representation formula is articulated in the subsequent (4):

$$x = \left[u, v, \gamma, h, \dot{u}, \dot{v}, \dot{\gamma}, \dot{h}\right]^T \quad (4)$$

The state vector x mathematically represents the state of a monitored object at a certain moment within the DeepSORT Kalman Filter. This vector comprises eight elements, each signifying the following:

u : The central horizontal coordinate (centre x-coordinate) of the object's enclosing box. This denotes the object's horizontal placement within the video frame.

v : The central vertical coordinate (central y-coordinate) of the object's enclosing box. This denotes the object's vertical placement within the video frame.

γ : The aspect ratio of the bounding box, determined by dividing its width by its height. This conveys details regarding the object's proportional configuration.

h : The elevation of the object's bounding box. This conveys details regarding the object's vertical dimension.

$\dot{u}$: The speed of the central horizontal coordinate. This assesses the velocity and direction of the horizontal movement of the bounding box's centre.

$\dot{v}$: The speed of the central vertical coordinate. This assesses the velocity and direction of the vertical movement of the bounding box's centre.

$\dot{v}$: The rate of variation in the aspect ratio. This assesses the temporal alterations in the object's proportionate form.

$\dot{h}$: The rate of alteration in the elevation of the bounding box. This assesses the temporal variation in the object's vertical dimensions..

^{T} : The notation at the vector's conclusion signifies that it is a transposed vector, represented in column format.

Appearance features are derived by extracting patches from the identified bounding box and processing them through a CNN architecture, such as ResNet-50 or MobileNet. A CNN produces a fixed-dimensional embedding vector, usually 128 dimensions that encapsulates the visual attributes of the object. Each track maintains a record of these embeddings, utilised to assess appearance similarity among frames.

In the DeepSORT algorithm, when numerous objects are detected in a video frame and we aim to track their identities across time, it is necessary to determine which object detection in the current frame corresponds to which object track from the preceding frame. The process of data matching, or association, is essential to prevent the misidentification of objects. DeepSORT employs two essential measures to facilitate precise association determinations: mahalanobis distance and cosine similarity [10].

Mahalanobis distance quantifies the proximity of the current object detection to the predicted position of the tracked object. In contrast to the conventional Euclidean distance, Mahalanobis distance incorporates uncertainty in the projected object position [31]. The covariance matrix obtained from the Kalman Filter employed in tracking signifies this uncertainty. Conversely, when the tracker exhibits high confidence in its forecast (shown by minimal covariance), a minor positional discrepancy will yield a substantial Mahalanobis distance. Conversely, if the tracker exhibits diminished confidence (high covariance), a greater positional discrepancy may still be seen as proximate. The Mahalanobis distance formula (5) between detection d_j, prediction y_i , utilising the covariance matrix S_i can be expressed as [32]:

$$d^{(1)}(i,j) = \left(d_j - y_i\right)^T S_i^{-1}\left(d_j - y_i\right) \qquad (5)$$

DeepSORT utilises both location and appearance information from each object. Each identified object has its visual characteristics retrieved via a deep neural network. These features are vectors that denote the object's distinctive visual attributes. Cosine Similarity is employed to assess the similarity of appearance features between the current detection and the appearance features of the object's trajectory recorded from prior frames. Cosine Similarity quantifies the angle between two feature vectors. When the angle is minimal (cosine value approaching 1), the two visual characteristics exhibit significant similarity, suggesting they are likely the same object. When the angle is substantial (cosine value approaching -1), the visual characteristics are markedly distinct.

The amalgamation of these two metrics generates an association cost matrix, which is resolved with the Hungarian Algorithm. To evaluate the visual resemblance between the bounding box of the currently tracked object (i) and the bounding box of the newly detected object (j), we employ an equation $d^{(2)}(i,j)$ (6) that computes the cosine distance, which is the inverse of the cosine similarity, between their appearance feature vectors $(r_j \ and \ r_i)$. The dot product $(r_j^T r_i)$ yields the cosine similarity value, and by subtracting it from 1, we derive the cosine distance. A small distance signifies maximal similarity in visual characteristics.

$$d^{(2)}(i,j) = min\left\{1 - r_j^T r_k^{(i)} \middle| r_k^{(i)} \epsilon R_i\right\} \qquad (6)$$

Upon establishing the association, the system classifies the track as matched, mismatched, or a novel detection. Upon a detection aligning with an existing track, the Kalman filter is revised, and the appearance features are recorded. In the absence of corresponding detections, the track persists for some frames prior to being eliminated. Conversely, novel, unassociated detections will generate new tracks if their confidence level is sufficiently high. Parameters, including maximum age, minimum hits, and matching threshold significantly influence the system's susceptibility to noise and visual disruptions.

D. Higher Order Tracking Accuracy (HOTA) Metric

Higher Order Tracking Accuracy (HOTA) is a principal evaluation statistic intended to furnish a singular, equitable assessment of numerous object tracking efficacy [6]. HOTA explicitly integrates accurate detection, proper association between detection and tracking, and exact localisation into a singular metric. A high HOTA is intuitively attained when the tracking system effectively identifies a significant percentage of target objects, persistently preserves their identities across frames, and accurately anticipates bounding boxes. HOTA is determined by the square root of the product of Detection Accuracy (DetA) and Association Accuracy (AssA), imposing a penalty for subpar performance in either dimension.

Detection Accuracy (DetA) is a sub-metric of HOTA that particularly evaluates the efficacy of the tracking system in identifying the proper item in each frame. DetA is computed by comparing the quantity of accurate detections (True Positives) to the total count of ground truth objects and resultant detections (True Positives + False Negatives + False Positives). In essence, DetA resembles the F1 score for detection tasks, balancing detection precision (minimising False Positives) and detection recall (minimising False Negatives). A high DetA signifies that the system can accurately identify the majority of target objects while minimising false detections.

Association Accuracy (AssA) is a sub-metric of HOTA that evaluates the precision with which a tracking system preserves the right object to identify throughout video frames. AssA evaluates the quality of the correlation between a detection in the current frame and the established track from the preceding frame. Calculating AssA entails contrasting the quantity of accurate associations with the entire number of potential linkages. An accurate association transpires when a detection is linked to the appropriate ground truth track. A high AssA signifies that the system can constantly monitor objects without frequent identity transitions or track fragmentation.

Localisation Accuracy (LocA) is a sub-metric of HOTA that assesses the precision with which the tracking system's bounding boxes identify the target item. LocA is generally computed as the mean Intersection over Union (IoU) between the predicted bounding box and the ground truth bounding box for all accurate detections. A high LocA signifies that the system not only accurately identifies items but also precisely forecasts their locations and dimensions.

False Positives (FP) refer to the detections produced by the tracking system that do not align with any actual ground truth objects. FP denotes the existence of an object that is, in reality, nonexistent (according to the ground truth annotation). A lower FP number indicates greater detection precision of the system.

False Negatives (FN) refer to the quantity of actual items that the tracking system failed to detect. FN denotes the system's inability to identify an object that was genuinely present in the frame. A lower FN value corresponds to an increased detection recall of the system.

Identity Precision (IDP) is a parameter that assesses the purity of the tracks produced by the tracking system. IDP is determined by the ratio of accurate associations to the total number of detections made by the system. A high IDP indicates that the majority of the generated tracks accurately reflect the genuine identify of the object.

Identity Recall (IDR) is a metric that evaluates the efficacy of a tracking system in preserving accurate ground truth identities. IDR is determined by the ratio of correct associations to the total count of ground truth items. A high IDR signifies that the system effectively monitors the majority of ground truth objects while preserving their identities.

Identity Switches (IDS) refer to the frequency with which a monitored object's identity is erroneously altered throughout a video sequence. A low IDS signifies enhanced stability and consistency in the performance of association for preserving object identities.

E. CLEAR Metric

The CLEAR (Classification of Events, Activities, and Relationships) metric serves as a benchmark for the quantitative and objective assessment of moving object tracker performance [33]. Multi-Object Tracking Accuracy (MOTA) serves as the principal composite statistic inside the CLEAR evaluation framework [34]. MOTA comprises three principal types of mistakes in tracking: false positives (FP), false negatives (FN), and identity switches (IDS) [33]. MOTA is formally computed as (7):

$$MOTA = 1 - \frac{|FN|+|FP|+|IDSW|}{gtDet} \quad (7)$$

In frame t, FN, FP, and IDSW denote the quantities of false negatives, false positives, and identity switches, respectively, whereas gt represents the count of ground truth objects in frame t. MOTA delivers a singular metric that evaluates both detection precision and identification coherence across time. An elevated MOTA value signifies superior tracking performance.

Multi-Object Tracking Precision (MOTP) assesses the accuracy of successfully localised tracked objects, excluding errors in detection or identity [35]. MOTP is determined by the average overlap, typically employing Intersection over Union (IoU), between the predicted bounding boxes and the ground truth bounding boxes for all true positives over all frames. MOTP can be formally expressed as follows (8) [5]:

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t C_t} \quad (8)$$

where, $d_{t,i}$ is the distance (e.g. 1−IoU) between detection i and the corresponding ground truth at frame t, and C_t is the number of correct matches at frame t. MOTP indicates the precision of the system in localising successfully tracked objects, with elevated values signifying superior localisation.

III. Method and Material

A. MOT Step

The implementation of a Multiple Object Tracking (MOT) system for road vehicles commences with the acquisition of an input video stream, sourced either directly from a camera or from pre-existing footage. The video is processed by decomposing it into a sequence of frame acquisitions, each of which is independently and sequentially processed. The last essential phase is object detection, which in this research was performed utilising different forms of the YOLOv8 architecture (nano, small, medium, large, or extra-large). The selection of the YOLOv8 variation considerably influences the equilibrium between inference speed and detection accuracy, which is the primary focus of this evaluation. The YOLOv8 detection results comprise detection output, a collection of bounding boxes around each identified vehicle, confidence ratings reflecting the amount of detection certainty, and class labels categorising the vehicle type.

Following detection, feature extraction is conducted for each identified object utilising a deep neural network to obtain appearance features, a crucial component of the DeepSORT algorithm. This pre-trained deep neural network produces Appearance Features, which are vector descriptors that distinctly characterise the visual attributes of each vehicle. Simultaneously, State Estimation and Prediction are executed for each existing object track utilising a Kalman Filter. This filter assesses the present object state (position, velocity, etc.) and forecasts the anticipated object states in the subsequent frame.

The subsequent phase is Association, which seeks to link the object detection in the present frame with the track prediction from the preceding frame. This procedure utilises the Hungarian Algorithm, which operates on a cost matrix. This cost matrix evaluates two primary metrics: distance and appearance are assessed based on the disparity in visual characteristics between the detection and the track, while the Mahalanobis Distance (Motion) quantifies the deviation between the detected position and the motion prediction, incorporating the uncertainty inherent in the Kalman filter prediction [36]. The Hungarian algorithm subsequently seeks the assignment with the lowest cost. Successfully linked detections are utilised for Track Updates, wherein the Kalman Filter of the track is revised using the most recent detection data, enhancing the estimation of the object's state. Ultimately, Track Management oversees the lifecycle of the object track through Initialisation for consistent new detections and Termination for tracks that remain undetected across multiple frames. The ultimate outcome of this method is Output: Tracked Objects, which displays the movement trajectory of each identified vehicle together with unique IDs that are preserved throughout the film. The phases of object tracking utilising Yolov8 and DeepSORT are illustrated in Fig. 1.

B. YOLOv8 Object Detection

The object detection procedure utilising YOLOv8 on a video stream commences with the acquisition of an input video stream, which may originate from a live roadside security camera or a recorded video. The video stream is segmented into a sequence of frame acquisitions, with each frame serving as a static visual analysis unit that will be processed sequentially by the YOLOv8 model. The subsequent stage is grid division, wherein each frame is partitioned into an S×S grid of cells [37]. This division seeks to localise items into designated regions of the frame, with each grid cell tasked with anticipating the object's centre contained within it.

YOLOv8 employs an anchor-free prediction methodology, distinguishing itself from its predecessors by explicitly forecasting the centre, height, and breadth of the bounding box in relation to each grid cell. This method offers enhanced adaptability in managing discrepancies in object dimensions and proportions. Each input frame undergoes a convolutional feature extraction process utilising the YOLOv8 CNN architecture [38]. This network systematically extracts visual information from low-level elements (edges and corners) to high-level components (object-specific shapes). YOLOv8 utilises an efficient backbone architecture, such as CSPDarknet [39] or the C2f module to achieve a balance between speed and accuracy.

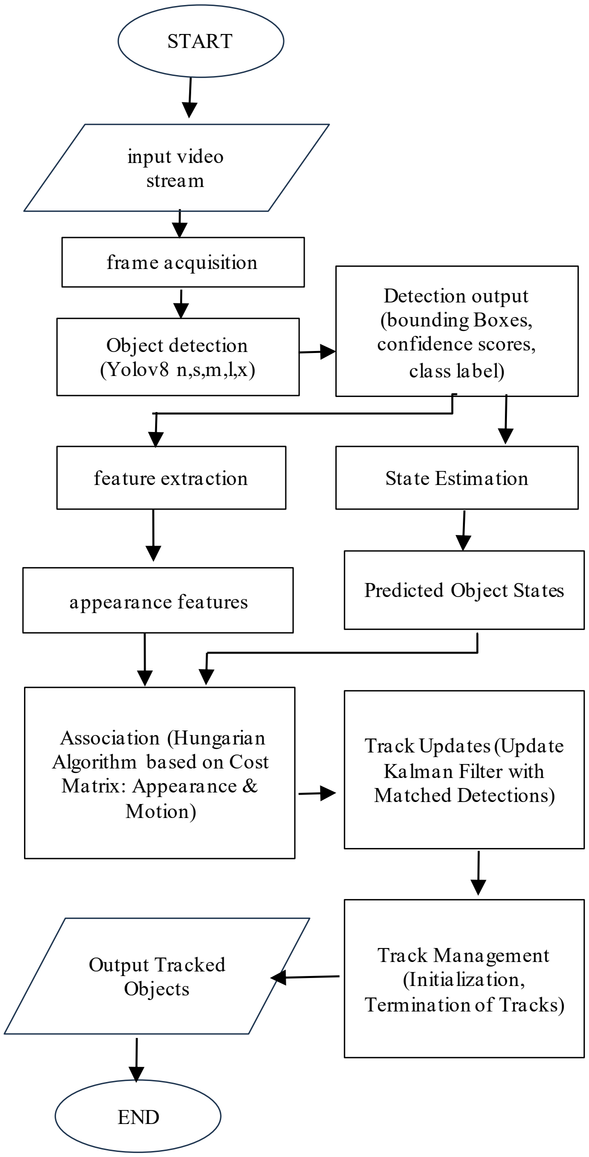
START
input video stream
frame acquisition
Object detection (Yolov8 n,s,m,l,x)
Detection output (bounding Boxes, confidence scores, class label)
feature extraction
State Estimation
appearance features
Predicted Object States
Association (Hungarian Algorithm based on Cost Matrix: Appearance & Motion)
Track Updates (Update Kalman Filter with Matched Detections)
Output Tracked Objects
Track Management (Initialization, Termination of Tracks)
END

Fig. 1. Multi object tracker step.

YOLOv8 concurrently produces a Prediction per Grid Cell for each grid cell, encompassing the coordinates (x,y,w,h) of the prospective bounding box, an Objectness Score reflecting the confidence level of the object within the bounding box, and Class Probabilities that denote the conditional probability for each object class upon detection. Due to the possibility of numerous detections of the same object from several grid cells or overlapping bounding box predictions, Non-Maximum Suppression (NMS) is utilised. This procedure iteratively identifies the bounding box with the highest confidence score and discards other bounding boxes that exhibit substantial overlap (measured by Intersection over Union - IoU) and possess lower confidence scores. This procedure persists until solely the most precise bounding box for each identified object is preserved. The culmination of this phase is the Detection Output, comprising a collection of filtered bounding boxes, each associated with a high confidence score and accompanying vehicle class label, which will subsequently function as input for the DeepSORT tracking phase.

C. Tracking Evaluation

Assessing the efficacy of a Multiple Object Tracking (MOT) system is essential for comprehending the proficiency of your algorithm in tracking things within video [5]. This process involves comparing the MOT system's output with manually annotated ground truth data. Metrics such as HOTA and CLEAR are employed to assess several dimensions of tracking precision.

The evaluation phase commences with Ground Truth Data (Manual Annotations). This constitutes the basis of the comprehensive assessment, featuring detailed manual annotations for each pertinent object (e.g. motorbikes, vehicles, buses, trucks) in every video frame. These annotations comprise precise bounding boxes and constant unique identifiers for each object throughout the video frames. Upon completion of the annotations, the MOT data must be prepared to ensure compatibility with the evaluation code. A standard format often comprises the frame number, object ID, and bounding box coordinates (x, y, width, height).

Subsequently, we present the Tracking Result Data (MOT System Output). This is the output produced by your MOT system, namely a synthesis of YOLOv8 and DeepSORT. This data includes frame-by-frame details regarding the identified and monitored objects, together with bounding boxes and IDs allocated by the tracking algorithm. Similar to the GT data, this tracking data must be prepared to adhere to the HOTA and CLEAR assessment requirements.

Upon formatting both data sets, the GT data and the tracking data, we proceed to the HOTA Evaluation phase. The HOTA code implementation will juxtapose the tracks produced by your system with the ground truth tracks for each object across the video sequence. The outcome of this assessment is HOTA Metrics, an extensive array of metrics that offers a thorough analysis of tracking performance. The metrics encompass: HOTA (which balances detection and association accuracy), DetA (detection accuracy), AssA (association accuracy), LocA (bounding box localisation accuracy), FP (false positives), FN (false negatives), IDP (identity precision), IDR (identity recall), and IDS (identity switches) [40], [41].

The CLEAR Metrics review is conducted concurrently with or subsequent to the HOTA review. GT data and tracking outcomes are used to compute these traditional metrics. The resulting CLEAR metrics encompass: MOTA (Multi-Object Tracking Accuracy), which integrates detection and association errors; MOTP (Multi-Object Tracking Precision), which assesses bounding box accuracy; MT (Mostly Tracked), the proportion of ground truth objects that are predominantly tracked; PT (Partially Tracked), the proportion of ground truth objects that are partially tracked; ML (Mostly Lost), the proportion of ground truth objects that are predominantly untracked; FP and FN (as in HOTA); IDS (number of identity switches); and Frag (Fragmentation), the frequency with which a track is interrupted and subsequently resumed.

The concluding phases involve the analysis and interpretation of HOTA results and CLEAR results. The metric data from both sets are assessed to comprehend the overall efficacy of the tracking system. The emphasis for HOTA is on the aggregate HOTA value, including the contributions of DetA and AssA. Analysing HOTA values across many YOLOv8 versions will identify the variation that provides optimal tracking performance. Alternative measurements offer supplementary insights into particular facets. Likewise, CLEAR metrics, particularly MOTA, offer a comprehensive summary, whereas MOTP emphasises bounding box precision. MT, PT, and ML measurements offer insights into the track quality of individual objects. The Mostly Tracked (MT) statistic quantifies the proportion of ground truth objects that are effectively monitored for the majority of their lifespan. An object is deemed "mostly tracked" if it is effectively linked to a track for a minimum of 80% of its overall duration in the video. MT offers insight into the system's capacity to sustain long-term monitoring of a singular item. Conversely, Mostly Lost (ML) quantifies the proportion of ground truth objects that are not monitored for the majority of their lifespan [31]. ML denotes the frequency with which the system fails to sustain long-term tracking of an item. Partially Tracked (PT) denotes the proportion of ground truth objects that do not belong to either the MT or ML categories. These things are monitored for a substantial duration of their existence, yet insufficiently to be categorised as MT or marginally enough to be designated as ML

False Positives (FP) refer to the aggregate number of detections produced by the tracking system that do not align with any actual objects in the complete video sequence. FP denotes erroneous detections, or backdrops erroneously identified as objects. False Negatives (FN) refer to the total count of actual objects that the tracking system fails to detect during the whole video sequence. FN denotes the system's inability to identify an object that is genuinely present. Identity Switches (IDS) refer to the cumulative instances in which a monitored object's identity is erroneously altered over the video sequence. A low IDS signifies enhanced identity consistency and the system's capacity to preserve individual object trajectories without interchanging them with those of other objects. Fragmentation (Frag) quantifies the frequency with which an object's trajectory is disrupted and subsequently re-established. Fragmentation transpires when an object is monitored, thereafter vanishes (lacking association with a detection in the subsequent frame), and later reemerges with the identical ID. Minimal fragmentation signifies a more stable and uninterrupted trajectory. A conclusion regarding Tracking Performance Comparison is reached through this analysis, pinpointing which YOLOv8 variants consistently yield the highest HOTA values and other pertinent metrics for tracking motorcycles, cars, buses, and trucks on highways, while also addressing the trade-off between accuracy and speed. The illustration depicting the phases of the object tracking evaluation procedure is presented in Fig. 2.

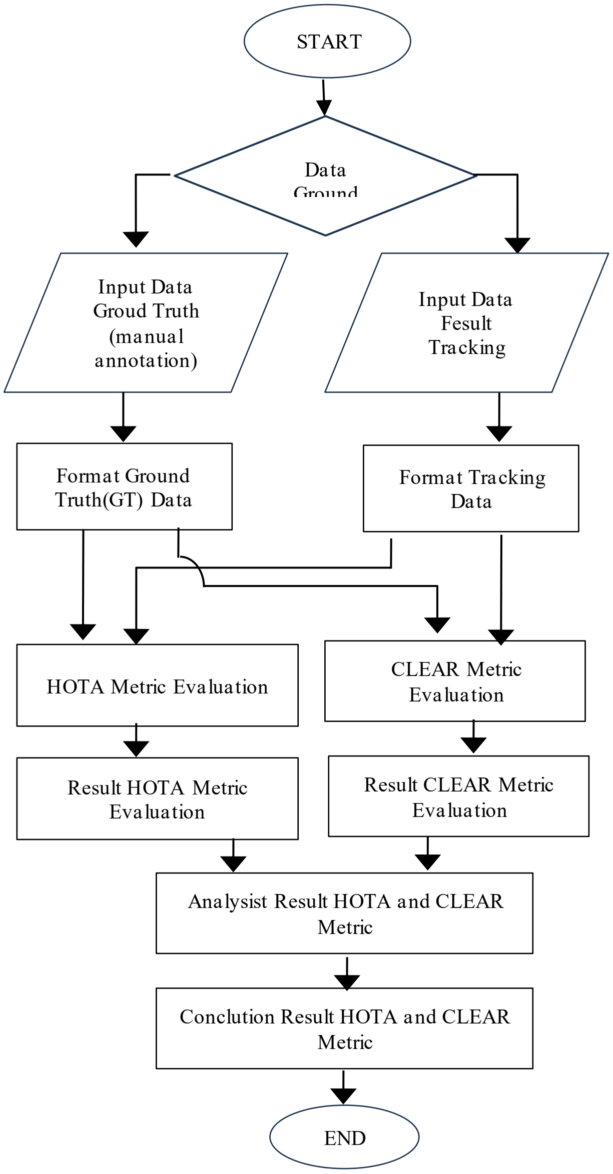
START
Data Ground
Input Data Groud Truth (manual annotation)
Input Data Fesult Tracking
Format Ground Truth(GT) Data
Format Tracking Data
HOTA Metric Evaluation
CLEAR Metric Evaluation
Result HOTA Metric Evaluation
Result CLEAR Metric Evaluation
Analysist Result HOTA and CLEAR Metric
Conclution Result HOTA and CLEAR Metric
END

Fig. 2. Object tracking evaluation process.

D. Material Research

This work employs a relevant and extensive video dataset for the analysis of vehicle object tracking. The video data included in this experiment was specifically obtained from CCTV security cameras operated by the Bali Provincial Transportation Agency (https://balisatudata.baliprov.go.id/ peta-cctv). This study employs a collection of 980 real-world traffic surveillance photos, featuring 8,829 identified items categorized into four classes: cars, motorcyclists, buses, and trucks. To ensure model robustness, data were collected from two types of road environments: straight roads and intersections, thereby capturing diverse vehicle dynamics and spatial configurations.

The dataset encompasses differences in traffic flow conditions—spanning free-flow, moderate, and congested traffic—allowing for the evaluation of the tracking model under different density levels. The video data collection period took place in September 2024. This period was chosen to represent typical traffic patterns in the region, omitting substantial fluctuations that may occur during vacations or special events. It is important to acknowledge that all data were collected under daytime conditions, so presenting nocturnal scenarios as a forthcoming issue for future research.

This video dataset predominantly comprises two primary categories: motorbikes and cars, which are the subject of this research. Bali possesses a substantial motorbike demographic that often engages with bigger four-wheeled vehicles, including cars, buses, and trucks. The varied dimensions, forms, and motion trajectories of these two vehicle categories pose intriguing problems for object tracking algorithms. This dataset was created to assess the efficacy of tracking algorithms in differentiating and maintaining the identity of diverse vehicle types in congested traffic conditions.

The gathered video data is subsequently utilised as input for the object tracking system created in this study, which incorporates the YOLOv8 object recognition framework and the DeepSORT tracking algorithm. Each video is segmented into a sequence of individual frames, which are subsequently analysed by the YOLOv8 model to identify the presence and location of cars. The YOLOv8 detector's output, comprising bounding boxes, confidence scores, and class labels for each identified vehicle, is subsequently fed into the DeepSORT algorithm. DeepSORT uses this information, in conjunction with visual appearance attributes derived from each detection and motion prediction via a Kalman filter, to temporally track objects while preserving the distinct identification of each vehicle across video frames.

The utilisation of authentic video footage from a congested traffic setting in Bali enhances the authenticity and practical significance of this study's findings. The variety of traffic circumstances and vehicle types in the dataset enables thorough evaluation of the tracking algorithm's performance in practical situations. Moreover, the use of data from public transportation infrastructure, specifically CCTV from the Department of Transportation, underscores the prospective implementation of this object tracking technology in forthcoming traffic management and transportation monitoring systems. The dataset's quality and attributes establish a crucial basis for assessing performance and recognising potential enhancements in the suggested object tracking methodology.

Moreover, the acquisition and utilisation of this video data were executed with appropriate consideration for pertinent ethical and privacy issues. The data utilised is publicly available and sourced from surveillance cameras deployed for traffic monitoring. This research does not entail the acquisition of personal data or information that could identify specific persons beyond the context of their automobiles on the roadway. The research primarily concentrates on the analysis and tracking of various vehicle types for scientific and technological advancement. This dataset is deemed representative and sufficient for assessing the efficacy of object tracking algorithms within the framework of intense and varied urban traffic.

IV. Result

A series of tests performed using 30-second video footage at a Bali crossroads yields significant visual insights into the evolutionary performance of several YOLOv8 model variants integrated with the DeepSORT algorithm. The quantitative assessment exclusively targets cars and motorbikes, the predominant vehicle categories at the site, while the visualisation additionally illustrates the system's efforts to identify other classes, including buses and trucks. Each graphic, depicting progressively intricate model variants (from n to x), clearly demonstrates the trade-off among speed, detection accuracy, and classification reliability, closely correlating with the data shown in the HOTA and CLEAR metrics tables. We executed the tests in succession, first with Yolov8 version n and progressing to version x. The identifying number (ID) of each object in every test was randomly and uniquely created. Subsequently, we juxtaposed the visualisation outcomes within the identical frame.

To assess the model's detection efficacy on our unique dataset, we computed the mean Average Precision metric at a threshold of 0.50 (mAP50) for all four specified vehicle categories. This metric quantitatively assesses the model's accuracy in object localisation, with a detection deemed correct if the overlap between the predicted box and the ground-truth box is above 50%. The visual outcomes of this accuracy evaluation for each category are elaborated in Fig. 3.

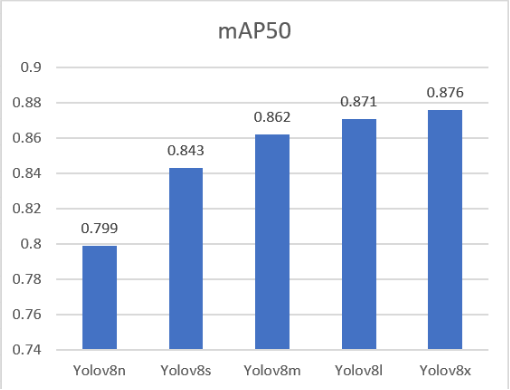

Fig. 3. mAP50 of Yolov8 variant.

Fig. 4 illustrates the outcomes of the YOLOv8n + DeepSORT model, highlighting the efficacy of the most lightweight and rapid variant. As a baseline model, its performance demonstrates notable shortcomings. The most significant inaccuracy was a false positive identification of object ID 415, erroneously classified as a "bus." In reality, this object was a massive billboard featuring a complex graphic, suggesting that the `n` model had difficulty differentiating between substantial vehicles and misleading background elements. Additionally, there were multiple false negatives, with roughly four motorcycles completely undetected. The inability to identify smaller, rapidly moving objects such as motorcycles is a prevalent deficiency in lighter models, which immediately resulted in diminished Detection Recall (DetRe) and MOTA scores in the quantitative assessment.

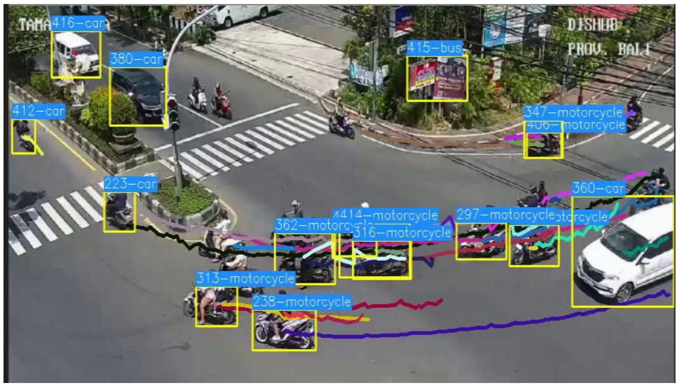

Fig. 4. Result from YOLOv8n + DeepSORT.

Fig. 5, which employs the YOLOv8s + DeepSORT model, shows that a progressive enhancement is observed. The count of undiscovered motorcycles decreases to three, signifying that the marginally more intricate 's' model had superior detection skills, hence diminishing the incidence of false negatives. Nonetheless, classification issues persist. Object ID 343, a white van, is erroneously categorised as a "bus". This categorisation of vehicles with unclear shapes (between a huge automobile and a small bus) is a persistent issue in the tests. Although these errors, when identified, do not directly influence detection metrics such as "DetA", they may affect evaluation if conducted purely on a per-class basis.

Fig. 6 presents the outcomes derived from the YOLOv8m combined with the DeepSORT model. There is a persistent misclassification, as the identical white van (now identified as 349) is once more erroneously categorised as a "bus". Nevertheless, it is evident that all motorcycles in the primary lane have been accurately detected, signifying an enhancement in detection recall relative to the 'n' and 's' variants. The enhancement in the capacity to identify all pertinent target objects is substantial and is evidenced by the increase in the "DetA" and "MOTA" scores in the metrics table. This suggests that with an increase in model size, its capacity to manage smaller and partially obscured objects becomes more resilient.

The trend of enhanced detection performance is notably illustrated in Fig. 7, produced by the YOLOv8l + DeepSORT model. An important discovery is that all vehicles, including motorcyclists, are accurately recognised and allocated bounding boxes. This is a notable accomplishment, indicating that the 'l' model possesses adequate representational capacity to tackle the detection issues at this congested intersection, successfully eradicating false negatives for this frame. Classification issues remain, as the white van (ID 290) is now erroneously categorised as a "truck". This underscores that enhancements in detection capabilities do not necessarily result in proportional advancements in classification accuracy for ambiguous instances.

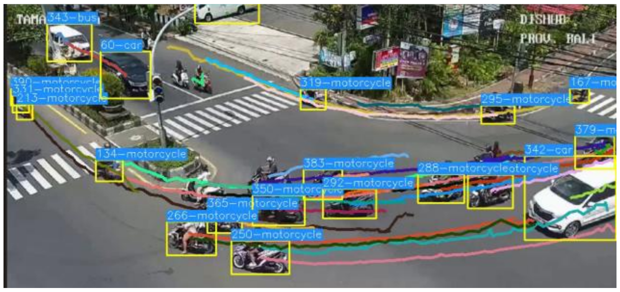

Fig. 5. Result from YOLOv8s + DeepSORT.

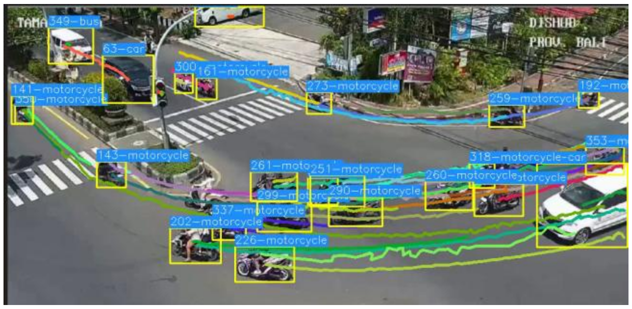

Fig. 6. Results from YOLOv8m + DeepSORT.

Ultimately, Fig. 8 illustrates the performance of the YOLOv8x + DeepSORT model, the most formidable option. Similar to the 'l' model, the 'x' model effectively identified all automobiles within the frame, demonstrating that the larger model markedly excels in recall performance. Nonetheless, it erroneously categorised a white van (ID 284) as a "truck". This indicates that, despite the model's extensive capacity, differentiating between visually analogous vehicle subclasses (such as vans, minibuses, and light trucks) continues to pose a significant challenge and may necessitate more varied training data or targeted fine-tuning methodologies. The visual progression from Fig. 4 to Fig. 8 clearly illustrates that augmenting the complexity of the YOLOv8 model enhances recall detection skills; yet, persistent challenges with ambiguous object classification underscore the necessity for future enhancement.

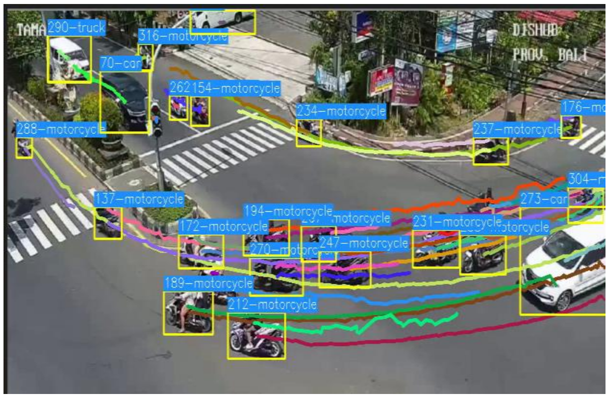

Fig. 7. Results from YOLOv8l + DeepSORT.

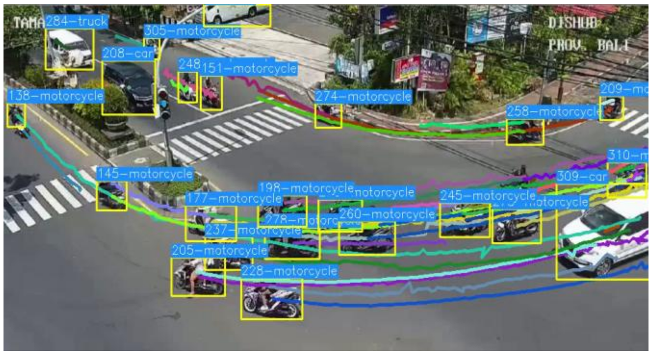

Fig. 8. Results from YOLOv8x + DeepSORT.

We subsequently present comprehensive HOTA evaluation results for the car class utilising a combination of the nano (n) and extra-large (x) YOLOv8 variants as detectors, paired with DeepSORT for association, and underscore the importance of these findings in determining the optimal detector architecture. The comprehensive assessment findings for all object classes and YOLOv8 versions will be displayed in tabular style and subsequently analysed in the following section. The CLEAR measure outcomes for this and further settings will be offered to offer a comprehensive view of performance tracking.

Tables I and II present the HOTA tracking outcomes for the “car” and “motorcycle” categories, respectively. According to the statistics in Tables I and II, a consistent trend is evident across all YOLOv8 model variants (from n to x): the overall tracking system exhibits marginally better and steadier performance for the car class than for the motorcycle class. The discrepancies in the primary HOTA scores are minimal; for instance, the top-performing x variant recorded a HOTA of 58.717, whereas the motorcycle attained 57.781. However, a thorough examination of the sub-metrics uncovers the origins of these performance variances and underscores the difficulties associated with motorcycle tracking.

TABLE I. HOTA TABLE FOR CAR CLASS

	Sub Metric	Yolov8 variant				
		n	*s*	*m*	*l*	*x*
	HOTA	53.137	55.541	55.596	58.477	58.717
SUB METRIC HOTA	DetA	47.491	50.915	52.408	55.602	54.479
	AssA	59.696	61.066	59.576	62.116	64.030
	DetRe	56.101	58.961	59.166	62.836	60.942
	DetPr	59.958	60.84	62.511	63.027	64.160
	AssRe	65.526	67.383	65.626	68.539	70.243
	AssPr	69.216	68.426	68.554	68.847	70.293
	LocA	74.073	73.390	73.131	73.298	73.746
	OWTA	57.869	59.970	59.304	62.415	62.379
	HOTA(0)	87.410	94.123	94.755	98.984	96.915
	LocA(0)	63.555	62.342	62.11	62.627	63.598
	HOTALocA(0)	55.554	58.678	58.853	61.991	61.636

TABLE II. HOTA TABLE FOR MOTOR CYCLE CLASS

		Yolov8 variant				
		n	*s*	*m*	*l*	*x*
	HOTA	51.545	54.140	55.372	57.703	57.781
SUB METRIC HOTA	DetA	47.820	48.684	53.002	55.624	53.842
	AssA	55.736	60.526	57.997	60.031	62.190
	DetRe	53.732	53.217	58.477	61.239	59.855
	DetPr	64.708	64.193	64.742	66.204	65.295
	AssRe	60.219	65.965	63.478	65.114	67.044
	AssPr	70.795	68.090	67.685	69.561	70.067
	LocA	75.339	73.128	73.620	74.365	74.317
	OWTA	54.656	56.696	58.198	60.611	60.993
	HOTA(0)	79.843	89.727	92.133	92.434	91.351
	LocA(0)	67.250	63.485	63.996	65.705	65.968
	HOTALocA(0)	53.694	56.963	58.962	60.734	60.262

Disaggregating HOTA results into Detection Accuracy (DetA) and Association Accuracy (AssA) provides an essential initial perspective. In these metrics, automobiles routinely surpass motorcycles. The performance disparity is more evident in AssA (e.g. 64,030 for cars against 62,190 for motorbikes in variation x) than in DetA (54,479 versus 53,842). This indicates that although modest difficulties exist in recognising motorcycles, a more substantial problem pertains to the tracker's capacity to sustain consistent motorbike IDs over time. The tracker misidentifies or erroneously exchanges motorbike IDs more frequently than it does for automobiles. Additional examination of the associated sub-metrics corroborates this conclusion. The primary distinction is evident in Association Recall (AssRe), with automobiles routinely achieving superior scores (e.g. 70,243 compared to 67,044 for variation x). Reduced AssRe scores for motorbikes immediately signify a track fragmentation problem. This indicates that a singular, ostensibly continuous motorbike route is often disrupted and recommenced as a new track with a distinct ID. This is probably because of the inherent characteristics of motorcycles: their reduced dimensions render them more vulnerable to complete obstruction by other vehicles, and their nimble and non-linear movement complicates motion prediction algorithms such as the Kalman Filter often employed in DeepSORT. Conversely, the AssPr (Association Precision) scores are almost indistinguishable across the two classes, suggesting that the challenge of consolidating several distinct objects into a single track (track merging) is not a significant differentiating factor.

The Detection Recall (DetRe) measure for motorbikes is marginally lower, indicating that the detector frequently overlooks motorcycles (false negatives) compared to vehicles, perhaps attributable to their reduced size. Notably, the Detection Precision (DetPr) for bikes is somewhat superior, suggesting that when the model identifies a motorbike, the forecast is generally more dependable (fewer false positives) compared to its predictions for cars. This HOTA measure research quantitatively demonstrates that the primary issue in tracking motorbikes, as opposed to cars, resides not merely in detection, but predominantly in data association, particularly in preserving trajectory consistency despite agile motions and frequent occlusions.

In summary, in the automobile class, the mean HOTA and sub-metrics rose from YOLOv8-n (61.3) to YOLOv8-x (65.7), exhibiting a standard variation of around 10 points, which signifies a constant enhancement in performance with larger models. In the motorbike class, a same tendency is observed: YOLOv8-n (61.3) enhances to YOLOv8-l/x (65.7); however, the elevated standard deviation suggests considerable variability among sub-metrics. Elevated maximum values (e.g. >90 for HOTA(0)) suggest that certain measurements demonstrate substantial performance improvements with larger YOLOv8 variations.

The CLEAR metric analysis offers a compelling narrative that seems contradicts the HOTA results. Table III is the result of CLEAR metric measurements for the car class and Table IV is the result of CLEAR metric measurements for the motorcycles class. Remarkably, the Multiple Object Tracking Accuracy (MOTA) measure consistently surpasses that of vehicles for motorcyclists across all iterations of the YOLOv8 model. In the most sophisticated model x, motorcycles attain a MOTA of 62,622, whereas vehicles achieve only 57,331. This discovery necessitates a thorough examination of the error components constituting the MOTA score to comprehend the fundamental dynamics.

The explanation for the elevated MOTA ratings for motorcycles is found in the analysis of the precision and recall metrics derived by CLEAR. The data indicate that tracking precision (CLR_Pr) for bikes markedly exceeds that of automobiles across all model variants (e.g. 84,199 for motorcycles compared to 80,179 for cars in variant x). MOTA is determined by the aggregate of errors, including False Positives (FP), False Negatives (FN), and ID Switches (IDS), and exhibits significant sensitivity to the quantities of FP and FN. Increased precision immediately indicates a markedly reduced quantity of. False Positives. Consequently, although motorbikes may encounter a higher frequency of association mistakes, a diminished number of false positives significantly decreases the overall errors in the MOTA formula, thereby enhancing their score relative to automobiles. A higher MOTA does not inherently indicate superior tracking performance; instead, it may signify the detector's exceptional accuracy for that specific class.

TABLE III. CLEAR METRIC TABLE FOR CAR CLASS

Sub Metric	Yolov8 variant				
	n	*s*	*m*	*l*	*x*
MOTA	44.415	47.904	50.888	56.604	57.331
MOTP	70.630	69.671	69.418	69.374	69.790
MODA	44.894	47.904	50.888	56.680	57.331
CLR_Re	69.231	72.408	72.768	78.188	76.158
CLR_Pr	73.990	74.715	76.882	78.426	80.179
MTR	19.737	28.571	31.746	41.935	34.722
PTR	48.684	35.714	20.635	29.032	45.833
MLR	31.579	35.714	47.619	29.032	19.444
IDSW	24	16	8	5	5
Frag	212	536	547	515	434

TABLE IV. CLEAR METRIC TABLE FOR MOTOR CYCLE CLASS

Sub Metric	Yolov8 variant				
	n	*s*	*m*	*l*	*x*
MOTA	53.906	51.935	55.943	63.886	62.622
MOTP	70.921	68.812	69.581	70.169	70.076
MODA	55.213	52.311	56.236	63.953	62.699
CLR_Re	69.125	67.606	73.279	78.227	77.183
CLR_Pr	83.246	81.550	81.131	84.569	84.199
MTR	31.169	44.495	28.571	30.435	40.370
PTR	58.442	37.615	44.643	58.261	42.609
MLR	10.390	17.890	26.786	11.304	16.522
IDSW	42	20	17	4	5
Frag	342	554	702	693	532

MOTA indicates a benefit for motorcycles; nevertheless, measures emphasising trajectory consistency present a contrasting narrative, aligning more closely with the prior HOTA findings. The quantity of FRAGs (Fragmentations) and IDSW (ID Switches) is continually elevated for motorcycles. In version x, motorcyclists exhibit 532 fragmentations, whereas vehicles have 434. In a less robust model, such as n, motorcyclists possess 42 IDSWs, whereas cars have just 24. These figures quantitatively validate that motorbike routes are more often fractured and their identities are more frequently exchanged. This substantiates the conclusion derived from the HOTA investigation that data association—sustaining consistent IDs amidst occlusion and dynamic movement, poses a far bigger problem for motorcycles.

Alternative trajectory quality measurements offer a more thorough perspective. The Mostly Tracked Ratio (MTR) for bikes is consistently superior (40,370 compared to 34,722 in the x variation); however, the Mostly Lost Ratio (MLR) is inferior (16,522 versus 19,444). This indicates that although motorcycle trajectories are prone to intermittent fragmentation, a greater percentage of the overall trajectories are effectively monitored during the majority of their duration. This suggests that your tracker is proficient at managing motorcycles in straightforward movement situations, although it struggles during complex events (sharp manoeuvres, significant occlusions) that result in IDSW and Frag. The nearly similar MOTP (Multiple Object Tracking Precision) ratings for both categories suggest that when an object is accurately recognised and appropriately associated, the bounding box localisation accuracy is equally proficient for both automobiles and motorcycles.

The results for motorbike tracking demonstrate that YOLOv8-x attains a HOTA score of 57.8, much surpassing the 51.5 achieved by YOLOv8-n, hence validating that larger detector variants enhance detection recall and association stability for tiny objects. Nonetheless, the aggregate HOTA values for motorbikes are inferior to those of cars, mostly attributable to recurrent occlusions, elevated object density, and the limited visual footprint of motorcycles in traffic scenarios. This finding aligns with the research of Frank Ngeni et al. [7], who indicated that tracking performance for small-scale and rapidly moving objects is susceptible to identification fragmentation, despite the utilisation of sophisticated detectors. Likewise, Jorge E. Espinosa1 et al. [4] noted that occlusion and swift manoeuvres markedly elevate ID switches in multi-object tracking benchmarks, underscoring the difficulties faced in this work. The comparisons indicate that although the integration of YOLOv8 with DeepSORT enhances motorcycle tracking relative to smaller detector variants, persistent challenges such as occlusion and identity consistency remain unaddressed, implying that future research should incorporate appearance-based re-identification modules or alternative tracking methods.

V. Conclusion

This article presents numerous major conclusions regarding the efficacy of a vehicle tracking system utilising the YOLOv8 detector in conjunction with the DeepSORT tracker, based on a thorough analysis undertaken. There exists a distinct positive correlation between the size of the YOLOv8 model and its detection accuracy; larger variants, such as YOLOv8l and YOLOv8x, markedly decrease the incidence of missed detections (false negatives), particularly for smaller objects like motorcycles, although difficulties in accurately classifying ambiguously shaped vehicles persist. Secondly, the dual-metric analysis provides significant complementary insights: the HOTA measure demonstrates that car tracking is more stable and consistent (higher "AssA" scores) compared to motorcycles, which frequently encounter track fragmentation due to their diminutive size, occlusion, and agile manoeuvres. Secondly, conversely, the CLEAR MOTA metric unexpectedly indicates elevated scores for motorcycles. This improvement is not attributable to enhanced association but rather to increased detection precision (fewer false positives), which disproportionately elevates the MOTA score and conceals deficiencies in tracking consistency (more ID shifts and fragmentation). This discovery significantly corroborates the original premise that a detection-biased statistic such as MOTA may yield an inadequate representation of actual tracking performance. This study indicates that HOTA offers a more equitable and comprehensive assessment by individually evaluating detection and association quality. This indicates that the selection of detector architecture and assessment metrics must be customised to the specific application requirements: in safety-critical systems, where identification consistency is essential, prioritising the HOTA association score is vital.

Subsequent study may focus on investigating the integration of YOLOv8 with different tracking algorithms beyond DeepSORT to enhance identification consistency and minimize trajectory fragmentation in multi-object tracking. A notable possibility is ByteTrack, which has a distinctive technique that accounts for low-rank detection boxes. Furthermore, advanced trackers like StrongSORT and OC-SORT merit assessment due of their enhancements in association consistency and tracking robustness in intricate traffic scenarios. Through the examination of diverse combinations, it is anticipated that a YOLOv8-based vehicle tracking system will attain enhanced performance and adaptability to a range of real-world conditions.

References

[1] M. Elassy, M. Al-Hattab, M. Takruri, and S. Badawi, "Intelligent transportation systems for sustainable smart cities," Transportation Engineering, vol. 16, p. 100252, Jun. 2024, doi: 10.1016/j.treng.2024.100252.

[2] L. Fei and B. Han, "Multi-Object Multi-Camera Tracking Based on Deep Learning for Intelligent Transportation: A Review," Sensors, vol. 23, no. 8, p. 3852, Apr. 2023, doi: 10.3390/s23083852.

[3] R. C. R. Nampalli, "Leveraging AI in Urban Traffic Management: Addressing Congestion and Traffic Flow with Intelligent Systems," Journal of Artificial Intelligence and Big Data, vol. 1, no. 1, pp. 86–99, May 2021, doi: 10.31586/jaibd.2021.1151.

[4] J. E. Espinosa, S. A. Velastin, and J. W. Branch, "Detection and Tracking of Motorcycles in Congested Urban Environments Using Deep Learning and Markov Decision Processes," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11524 LNCS, Springer Verlag, 2019, pp. 139–148. doi: 10.1007/978-3-030-21077-9_13.

[5] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," EURASIP J Image Video Process, vol. 2008, pp. 1–10, 2008, doi: 10.1155/2008/246309.

[6] J. Luiten et al., "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking," Int J Comput Vis, vol. 129, no. 2, pp. 548–578, Feb. 2021, doi: 10.1007/s11263-020-01375-2.

[7] F. Ngeni, J. Mwakalonge, and S. Siuhi, "Solving traffic data occlusion problems in computer vision algorithms using DeepSORT and quantum computing," Journal of Traffic and Transportation Engineering (English Edition), vol. 11, no. 1, pp. 1–15, Feb. 2024, doi: 10.1016/j.jtte.2023.05.006.

[8] X. Hou, Y. Wang, and L.-P. Chau, "Vehicle Tracking Using Deep SORT with Low Confidence Track Filtering," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, Sep. 2019, pp. 1–6. doi: 10.1109/AVSS.2019.8909903.

[9] T.-N. Doan and M.-T. Truong, "Real-time vehicle detection and counting based on YOLO and DeepSORT," in 2020 12th International Conference on Knowledge and Systems Engineering (KSE), IEEE, Nov. 2020, pp. 67–72. doi: 10.1109/KSE50997.2020.9287483.

[10] L. Lin, H. He, Z. Xu, and D. Wu, "Realtime Vehicle Tracking Method Based on YOLOv5 + DeepSORT," Comput Intell Neurosci, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/7974201.

[11] D. Guo, Z. Li, H. Shuai, and F. Zhou, "Multi-Target Vehicle Tracking Algorithm Based on Improved DeepSORT," Sensors, vol. 24, no. 21, p. 7014, Oct. 2024, doi: 10.3390/s24217014.

[12] O. Angah and A. Y. Chen, "Tracking multiple construction workers through deep learning and the gradient based method with re-matching based on multi-object tracking accuracy," Autom Constr, vol. 119, p. 103308, Nov. 2020, doi: 10.1016/j.autcon.2020.103308.

[13] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, and H. Lu, "Improving Multiple Object Tracking with Single Object Tracking," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2021, pp. 2453–2462. doi: 10.1109/CVPR46437.2021.00248.

[14] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," Artif Intell, vol. 293, p. 103448, Apr. 2021, doi: 10.1016/j.artint.2020.103448.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 6517–6525. doi: 10.1109/CVPR.2017.690.

[17] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.02767

[18] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.10934

[19] B. Mahaur and K. K. Mishra, "Small-object detection based on YOLOv5 in autonomous driving systems," Pattern Recognit Lett, vol. 168, pp. 115–122, 2023, doi: https://doi.org/10.1016/j.patrec.2023.03.009.

[20] C. Li et al., "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications," Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.02976

[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2023, pp. 7464–7475. doi: 10.1109/CVPR52729.2023.00721.

[22] M. Sohan, T. Sai Ram, and Ch. V. Rami Reddy, "A Review on YOLOv8 and Its Advancements," in Data Intelligence and Cognitive Informatics, S. and F.-G. P. Jacob I. Jeena and Piramuthu, Ed., Singapore: Springer Nature Singapore, 2024, pp. 529–545. doi: 10.1007/978-981-99-7962-2_39.

[23] M. Safaldin, N. Zaghden, and M. Mejdoub, "An Improved YOLOv8 to Detect Moving Objects," IEEE Access, vol. 12, pp. 59782–59806, 2024, doi: 10.1109/ACCESS.2024.3393835.

[24] I. N. E. Indrayana, M. Sudarma, I. K. G. D. Putra, and A. A. K. O. Sudana, "Improve Nighttime Highway Vehicles and Pedestrian Detection Using Yolov8+CLAHE," in 2024 Ninth International Conference on Informatics and Computing (ICIC), IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ICIC64337.2024.10956682.

[25] I. Purwita Sary, E. Ucok Armin, S. Andromeda, E. Engineering, and U. Singaperbangsa Karawang, "Performance Comparison of YOLOv5 and YOLOv8 Architectures in Human Detection Using Aerial Images," Ultima Computing : Jurnal Sistem Komputer, vol. 15, no. 1, 2023.

[26] H. K. Jooshin, M. Nangir, and H. Seyedarabi, " Inception ‐ YOLO: Computational cost and accuracy improvement of the YOLOv5 model based on employing modified CSP, SPPF, and inception modules, " IET Image Process, vol. 18, no. 8, pp. 1985 – 1999, Jun. 2024, doi: 10.1049/ipr2.13077.

[27] A. N. Alhawsawi, S. D. Khan, and F. U. Rehman, "Enhanced YOLOv8-Based Model with Context Enrichment Module for Crowd Counting in Complex Drone Imagery," Remote Sens (Basel), vol. 16, no. 22, p. 4175, Nov. 2024, doi: 10.3390/rs16224175.

[28] W. Yang, X. Tang, K. Jiang, Y. Fu, and X. Zhang, "An Improved YOLOv5 Algorithm for Vulnerable Road User Detection," Sensors, vol. 23, no. 18, p. 7761, Sep. 2023, doi: 10.3390/s23187761.

[29] M. I. H. Azhar, F. H. K. Zaman, N. Md. Tahir, and H. Hashim, "People Tracking System Using DeepSORT," in 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), IEEE, Aug. 2020, pp. 137–141. doi: 10.1109/ICCSCE50387.2020.9204956.

[30] D. M. Jiménez-Bravo, Á. Lozano Murciego, A. Sales Mendes, H. Sánchez San Blás, and J. Bajo, "Multi-object tracking in traffic environments: A systematic literature review," Neurocomputing, vol. 494, pp. 43–55, Jul. 2022, doi: 10.1016/j.neucom.2022.04.087.

[31] E. Bayraktar, "Advanced Kalman Filter Optimization for Efficient Multi-Object Tracking in Computer Vision," in 2024 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ASYU62119.2024.10757018.

[32] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in 2017 IEEE International Conference on Image Processing (ICIP), IEEE, Sep. 2017, pp. 3645–3649. doi: 10.1109/ICIP.2017.8296962.

[33] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," EURASIP J Image Video Process, vol. 2008, pp. 1–10, 2008, doi: 10.1155/2008/246309.

[34] A. B. Holla, M. M. M. Pai, U. Verma, and R. M. Pai, "Vehicle Re-Identification and Tracking: Algorithmic Approach, Challenges and Future Directions," IEEE Open Journal of Intelligent Transportation Systems, vol. 6, pp. 155–183, 2025, doi: 10.1109/OJITS.2025.3538037.

[35] L. Sharan, H. Kelm, G. Romano, M. Karck, R. De Simone, and S. Engelhardt, "mvHOTA: A multi-view higher order tracking accuracy metric to measure temporal and spatial associations in multi-point tracking," Comput Methods Biomech Biomed Eng Imaging Vis, vol. 11, no. 4, pp. 1281–1289, Jul. 2023, doi: 10.1080/21681163.2022.2159535.

[36] V. Nayak, B. Sonar, N. Marali, K. Mallibhat, and P. Nissimagoudar, "Comparative Analysis of Neural Trajectory Prediction Models for Vehicle Tracking," in Proceedings of 5th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, J. M. Gunjan Vinit Kumar and Zurada, Ed., Singapore: Springer Nature Singapore, 2025, pp. 37–50. doi: 10.1007/978-981-97-8865-1_4.

[37] E. Syahrudin, E. Utami, and A. D. Hartanto, "Object Detection with YOLOv8 and Enhanced Distance Estimation Using OpenCV for Visually Impaired Accessibility," JOIV : International Journal on Informatics Visualization, vol. 9, no. 2, p. 575, Mar. 2025, doi: 10.62527/joiv.9.2.2826.

[38] S. Pudaruth, I. M. Boodhun, and C. W. Onn, "Reducing Traffic Congestion Using Real-Time Traffic Monitoring with YOLOv8," International Journal of Advanced Computer Science and Applications, vol. 15, no. 10, 2024, doi: 10.14569/IJACSA.2024.01510109.

[39] R. Kridalukmana, D. Eridani, R. Septiana, and I. P. Windasari, "YoloV8, EfficientNetv2, and CSP Darknet Comparison as Recognition Model's Backbone for Drone-Captured Images," JOIV : International Journal on Informatics Visualization, vol. 9, no. 2, p. 683, Mar. 2025, doi: 10.62527/joiv.9.2.2880.

[40] Y. Li et al., "Beyond MOT: Semantic Multi-object Tracking," 2025, pp. 276–293. doi: 10.1007/978-3-031-72761-0_16.

[41] S. V. Ganesh, Y. Wu, G. Liu, R. Kompella, and L. Liu, "Amplifying Object Tracking Performance on Edge Devices," in 2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI), IEEE, Nov. 2023, pp. 83–92. doi: 10.1109/CogMI58952.2023.00021.