# Enhancing Multi-Object Tracking with ByteTrack, YOLO, and SAM: Integrating Detection and Segmentation for Improved Accuracy and Efficiency

Xingjian Gao, Kaicheng Yang, Yuqi Wang
University of Alberta
Edmonton, AB, Canada
{xg6,ky1,yuqi17}@ualberta.ca
https://github.com/prometheusowner/MOT-ByteTrack-Tracking-and-Prediction.git

## Abstract

*We introduce our System for Traffic Prediction, a sophisticated system employing advanced computer vision techniques to forecast the trajectories of various objects in complex traffic scenarios. This system integrates comprehensive approaches for both anticipatory action prediction and real-time monitoring across diverse traffic participants, including cars, pedestrians, and cyclists. By leveraging state-of-the-art object detection and tracking technologies, our system facilitates the monitoring of traffic dynamics, enabling preemptive measures by authorities. This not only enhances urban safety but also boosts the efficiency of traffic management.*

## 1. Introduction

The advent of computer vision (CV) technology has dramatically transformed various industries, facilitating the automation and enhancement of visual task processing. This technology fundamentally relies on the precise detection, recognition, and tracking of objects within images and video data. These capabilities are critical for a wide array of applications, including autonomous driving, surveillance, medical imaging, and retail. As the applications of computer vision broaden, the development of sophisticated systems capable of interpreting complex visual environments becomes increasingly crucial. This evolution underscores the importance of advancing computer vision technologies to meet the growing demands for accuracy and efficiency in real-world applications.

## 2. Literature Review

### 2.1. CV labeling

CV labeling is essential in developing and refining computer vision systems, enabling machines to interpret and understand visual data [5, 10]. This process involves annotating images or videos with descriptive labels that detail the visual content, ranging from object identification to complex interactions between entities.

The accuracy and quality of these annotations are crucial for the performance of computer vision models, especially in precision-dependent tasks like object detection, segmentation, and tracking. As computer vision technology proliferates across various domains—from autonomous vehicles to medical diagnostics—the need for robust, efficient, and precise CV labeling processes has become more pronounced.

CV labeling faces multiple challenges, including the requirement for high-quality annotations that reflect the complexity of real-world scenes, scalability for large datasets, and adaptability to changing needs. The labeling process is also typically labor-intensive and susceptible to human error, driving the need for enhanced automation and better tooling.

In response, recent advancements in CV labeling have incorporated machine learning techniques to automate parts of the labeling process, thereby improving both efficiency and accuracy. Additionally, the rise of crowdsourcing platforms has revolutionized data annotation, enabling the rapid and diverse expansion of annotated datasets.

This paper delves into the pivotal role of advanced labeling techniques in boosting the accuracy and efficiency of multi-object tracking systems, particularly through the lens of ByteTrack. ByteTrack employs sophisticated labeling strategies to enhance tracking performance in dynamically complex environments, demonstrating the transformative impact of innovative labeling solutions in contemporary computer vision applications.

### 2.2. Segment Anything Model (SAM)

SAM is a cutting-edge approach to data segmentation and analysis [6]. It presents a paradigm change in how com-

panies handle the categorization and analysis of data from a variety of sources. SAM, at its heart, uses complex algorithms to divide and classify information across a variety of domains, such as market trends and consumer behavior. This creative method enables companies to obtain a thorough grasp of their data environment, providing insightful information that facilitates well-informed decision-making.

The purposeful application of cutting-edge statistical and machine learning techniques is what makes SAM what it is. Through the use of these potent tools, SAM may reveal complex patterns and correlations that are hidden inside datasets. This in-depth degree of research gives companies substantial insights that go beyond cursory observations, enabling them to find untapped possibilities and reduce possible hazards [6, 12].

Graph-based Segmentation divides images into disjoint sets; Point-to-Surface Transition detects boundaries; region growing expands regions from seed points; boundary detection identifies edges; and hierarchical clustering clusters pixels according to similarity. These techniques are some of the methods used in image segmentation. These varied techniques provide a range of segmentation strategies, each with special advantages and uses.

In General, SAM is a revolutionary method of segmenting and analyzing data that is redefining how companies extract value from their data assets. SAM helps companies find hidden patterns and correlations in their data, facilitating informed decision-making and opening up fresh avenues for development and innovation. It does this by utilizing complex algorithms and cutting-edge techniques. SAM shows itself as a potent tool for companies looking to prosper in a more data-driven and competitive world because to its adaptability and agility.

### 2.3. You Only Look Once (YOLO)

YOLO is a popular computer vision technique, it uses a single neural network pass for object recognition and classification [4, 9].

In 2020, Ultralytics introduced YOLOv5, a departure from the traditional YOLO series. It comes with a redesigned architecture featuring Cross-Stage Partial connections (CSP) and Path Aggregation Network (PANet) for better performance. YOLOv5 offers variants like YOLOv5s to YOLOv5x, allowing users to prioritize either speed or accuracy based on their requirements. It delivers impressive performance, almost reaching the efficiency of EfficientDet AP, while also achieving higher Frames Per Second (FPS) [3, 4].

YOLOv8, the latest iteration from Ultralytics, stands as the cutting-edge model in the YOLO series (Fig.1). Building upon the success of its predecessors, YOLOv8 introduces new features and enhancements for improved performance, flexibility, and efficiency. It offers different sizes from YOLOv8s, to YOLOv8x. YOLOv8 supports a wide range of vision AI tasks, including detection, segmentation, pose estimation, tracking, and classification, making it versatile across various applications and domains [3].

In conclusion, YOLO and its variants represent a significant advancement in computer vision, altering the way humans identify things in pictures. They're helpful in numerous contexts thanks to their innovative techniques and constant advancements. As YOLO improves, it opens up new possibilities for image analysis, allowing individuals and companies to employ computer vision in more inventive ways.

### 2.4. Multi-Object Tracking (MOT)

MOT is a critical component of CV, essential for the detection and tracking of multiple objects across sequential frames. This technology is integral to a variety of applications, including surveillance systems, autonomous driving, and activity recognition. In these settings, comprehending the dynamics of multiple interacting entities is crucial [7, 11].

The efficacy of MOT systems is paramount for the precise and instantaneous interpretation of dynamic scenes. Conventional MOT methodologies often encounter substantial hurdles, such as elevated computational requirements, complex data association tasks, and managing occlusions where objects are temporarily obscured. These challenges are primarily due to the necessity of maintaining consistent object identities amid changing perspectives, lighting conditions, and environmental obstructions.

The performance of MOT systems is commonly evaluated based on their capability to minimize identity switches and tracking fragmentations, as well as their resilience in managing scenes with a diverse and complex array of objects [8]. With the growing real-world demands for MOT technology, advancing more sophisticated and efficient tracking algorithms continues to be a critical area of focus within computer vision research.

### 2.5. ByteTrack

Developed by Zhang [13], ByteTrack marks a significant progression in the MOT domain, primarily through its innovative use of deep learning techniques to enhance tracking robustness and efficiency. ByteTrack confronts the limitations of traditional tracking methods by employing a simple yet effective strategy for associating high-confidence detections across video frames.

Central to ByteTrack's approach is the emphasis on high-confidence detections to maintain consistent object identities over time, significantly mitigating the prevalent issue of identity switches, where two objects are mistakenly swapped. Additionally, ByteTrack integrates both short-term and long-term appearance cues to adeptly manage oc-

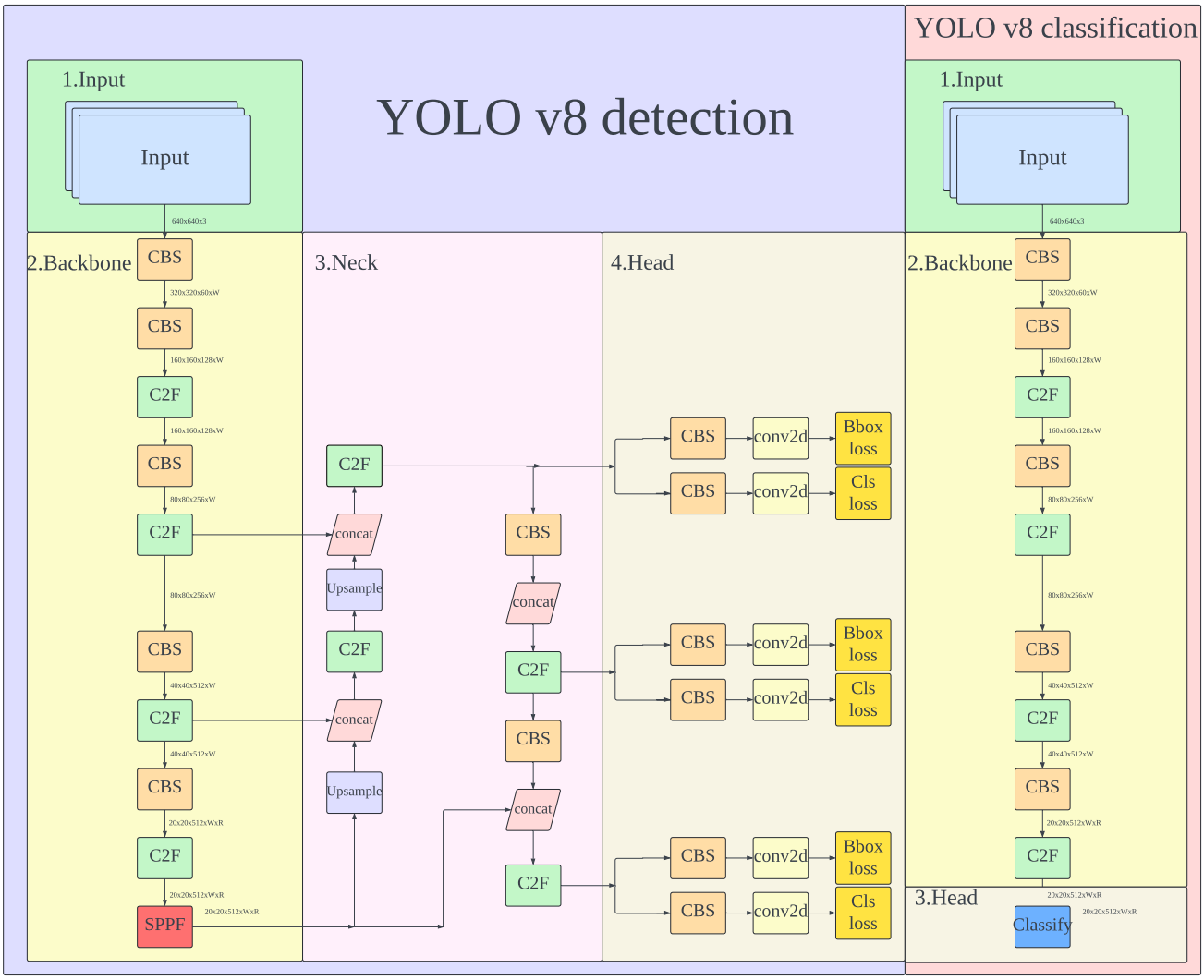


Figure 1. Flowchart illustrating the architecture of the YOLO model, detailing its structural components and data processing pathways.

clusions and complex interactions among objects.

The efficiency of ByteTrack is also elevated by its streamlined model architecture, designed to deliver real-time performance with minimal computational demand [13]. This attribute renders it especially advantageous for applications like autonomous driving, where rapid and accurate video data processing is crucial.

Furthermore, ByteTrack has demonstrated exceptional performance on standard MOT benchmarks such as MOT17 and MOT20, surpassing existing state-of-the-art models in both accuracy and efficiency. This success not only highlights ByteTrack's capability but also positions it as a transformative tool for researchers and practitioners in the computer vision community, offering unprecedented performance levels in practical tracking scenarios.

## 3. The Proposed Method/Pipeline

Prediction and Analysis of Traffic. In recent years, there has seen tremendous progress in the field of traffic analysis and prediction, mostly because of the growing need for improved effective urban mobility solutions. Deterministic models, which tend to provide averaged and inaccurate forecasts, were frequently used in traditional approaches to traffic analysis. This constraint has been overcome because of the advent of stochastic models. Furthermore, it has been demonstrated that using diffusion models can enhance the precision and effectiveness of traffic forecast algorithms. Object identification algorithms, YOLO and SAM, have advanced recently and improved multi-tracking systems' real-time traffic analysis capabilities. The basis for our proposed Multi-Tracking System is laid by these algorithms,

and it uses deep learning techniques to reliably recognize and track automobiles, pedestrians, and bikes in complicated traffic scenarios.

Development of Tracking Systems. Our Tracking System intends to accomplish full traffic analysis and prediction capabilities, building upon the advances in object identification and tracking algorithms. Motivated by the notable achievements of deep learning-based methodologies in other fields, our system employs convolutional neural networks (CNNs) [2] to effectively process and evaluate video streams instantaneously. To make sure precise object trajectories over time, we use sophisticated tracking algorithms, such as Kalman filters. We include cutting-edge algorithms like SAM and YOLO, which provide strong performance in recognizing and tracking objects in dynamic traffic settings, to further improve the accuracy of object recognition.

### 3.1. Kalman Filter

The Kalman Filter, a cornerstone in predictive algorithms, is widely utilized in MOT for its effectiveness in managing linear systems subject to Gaussian noise [1]. It is particularly adept at providing continuous estimation and updates of object states—like position and velocity—based on sequential measurements over time.

In ByteTrack, the Kalman Filter has been innovatively adapted not only to predict but also to refine the states of objects using incoming observational data, thereby elevating the tracker's precision and dependability [13]. ByteTrack employs the Kalman Filter to adeptly forecast the future state of each object between video frames. This application ensures fluid tracking continuity, even when objects are occluded or momentarily undetectable, thus minimizing typical issues such as identity switches and track fragmentation.

Furthermore, the integration of the Kalman Filter into ByteTrack highlights its efficiency in real-time operations. It enables ByteTrack to perform rapid state estimation and correction, essential for handling dynamic scenes in applications like autonomous driving and urban surveillance. By leveraging this established yet robust method, ByteTrack significantly advances traditional MOT techniques, offering superior performance with decreased computational demands.

The innovative application of the Kalman Filter in ByteTrack not only underscores its practical value in addressing contemporary tracking challenges but also marks a notable evolution from its conventional usage, demonstrating the filter's versatility and sustained significance in sophisticated computer vision applications.

### 3.2. Movement Prediction Using Historical Frame Data

Expanding upon the foundational principles of the Kalman Filter, this section introduces an innovative movement prediction approach within MOT that utilizes historical frame data to enhance predictive accuracy. By incorporating information from the preceding 10 frames, this method enriches the Kalman Filter's predictive capabilities by adding comprehensive temporal context [13].

This advanced movement prediction technique initializes the state estimation of the tracker with data derived from previous frames. By analyzing trajectories and velocities calculated from these historical frames, the tracker constructs a more precise prediction model for the future positions of objects. This improvement is particularly vital in environments with erratic or obscured object motions, ensuring continuity and reliability in the tracking process.

Employing a 10-frame historical window provides a nuanced understanding of an object's dynamics, crucial for anticipating sudden directional or speed changes. Such capability proves invaluable in complex scenarios like urban traffic systems or crowded public spaces, where accurate movement prediction is essential for safety and operational efficiency.

Furthermore, this approach not only augments the traditional functionality of the Kalman Filter by adding depth to the prediction process but also shifts the system from a reactive to a proactive stance. It enables the forecast of future states based on accumulated historical data, a key enhancement for advanced tracking applications where foresight can dramatically improve system effectiveness.

Integrating these historical insights into the predictive model not only boosts accuracy but also enhances robustness against temporary occlusions and detection failures. This strategy represents a significant evolution in MOT technologies, effectively bridging the gap between traditional tracking methodologies and the requirements of intricate, dynamic tracking environments.

### 3.3. Background Generation

In our methodology, we utilize the SAM to generate a clean background frame that excludes all moving objects. This is achieved through SAM's advanced segmentation capabilities, which identify and remove objects from video frames. Subsequently, these gaps are filled with contextually appropriate background details. This critical step prepares the scene for the subsequent reintroduction of objects at their predicted locations. By generating a background devoid of dynamic elements, we ensure that the further steps of object manipulation rely on a stable and uncontaminated environment, crucial for accurate predictions.

### 3.4. Object Isolation and Simulation

With a clear background established, the next phase involves isolating the moving objects from their original frames and simulating their future movements. This process leverages historical data from the last 10 frames to predict the objects' trajectories using a modified Kalman Filter. This adaptation not only considers the objects' current motion but also employs predictive models to forecast their future positions. The isolated objects are then digitally reintegrated into the clean background frame at these new predicted locations. This technique allows for precise control over each object's placement, facilitating the simulation of potential future frames where objects have shifted according to their projected paths.

### 3.5. Enhancing Tracking Accuracy

The primary objective of our methodology is to enhance the accuracy and efficiency of the tracking system. By generating prediction frames where objects are independently manipulated within a static background, our system is better equipped to predict movements and interactions, such as potential collisions or overlaps in densely populated scenes. This approach proves particularly advantageous in complex environments where traditional tracking methods might struggle due to occlusions or closely interacting objects. Additionally, this technique alleviates the computational load during live operations by addressing the most intensive processing tasks upfront. This preparation allows for faster and more precise real-time tracking and forecasting, an invaluable asset in applications demanding high predictive accuracy, such as autonomous vehicle navigation and proactive security monitoring systems.

## 4. Empirical Experiments and Analyses

### 4.1. Dataset Introduction and Preprocessing

In this work, we focus on introducing the datasets necessary for building an efficient multi-object detection, counting, and tracking system. Choosing high-quality datasets and performing appropriate preprocessing are crucial for training deep learning models, as these factors play a decisive role in model performance and generalization capabilities. Our dataset includes 5,542 high-quality images, covering various scenes to provide a rich background for both training and testing the model. Specifically, the dataset comprises 2,856 training images, 1,343 validation images, and 1,343 test images, allowing the model to train on diverse data while effectively assessing the model's generalization capabilities through the validation and test sets. This distribution ensures the comprehensiveness of the dataset in all aspects, providing a solid foundation for comprehensive model training (Fig. 2).

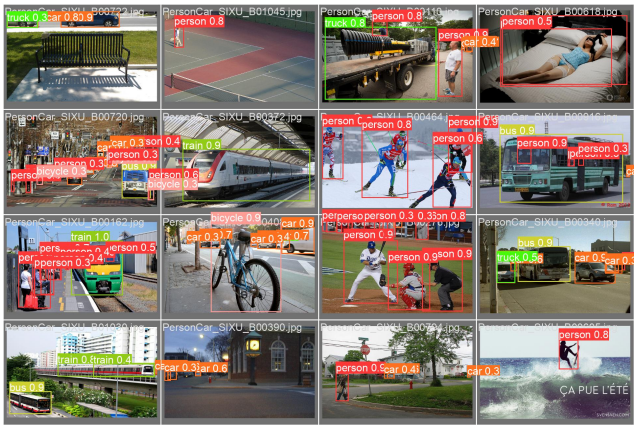

Figure 2. Labeled Images by trained YOLOv8 model, the model demonstrates powerful detection capabilities and identifies the confidence and category of the detected targets.

During preprocessing, all images undergo automatic orientation correction to eliminate inconsistencies due to different shooting angles of devices. Moreover, by removing EXIF data, we streamline the image data to enhance the efficiency of subsequent processing. Then, to meet the model's input requirements, all images are resized to 416x416 pixels. Although this resizing method may cause some distortion in image proportions, it provides uniformity for the model's input, which helps simplify the model architecture and improve computational speed.

### 4.2. Dataset Distribution and Impact on Model Training

The distribution of categories in the dataset is crucial, as it directly affects the model's classification performance. Based on the provided dataset distribution images (Fig. 4), we observe that the "person" category far exceeds other categories, such as "bicycle," "car," and "motorcycle." This imbalanced distribution suggests that the model might focus more on detecting the "person" category due to the abundance of samples available for learning. However, specific strategies like data augmentation or resampling techniques can be employed during training to prevent overfitting to the "person" category, while ensuring that other categories are also accurately detected.

Additionally, the distribution images also show a heatmap of target locations in the images and the distribution of target sizes. We can see that the distribution of target locations is relatively uniform, but the density is higher in the central area, possibly due to the positioning of the camera equipment and the tendency of targets to be located in the center of the field of view. The width and height distribution graphs of the targets reveal the diversity of target sizes, which is essential for training models capable of recognizing targets of different scales.

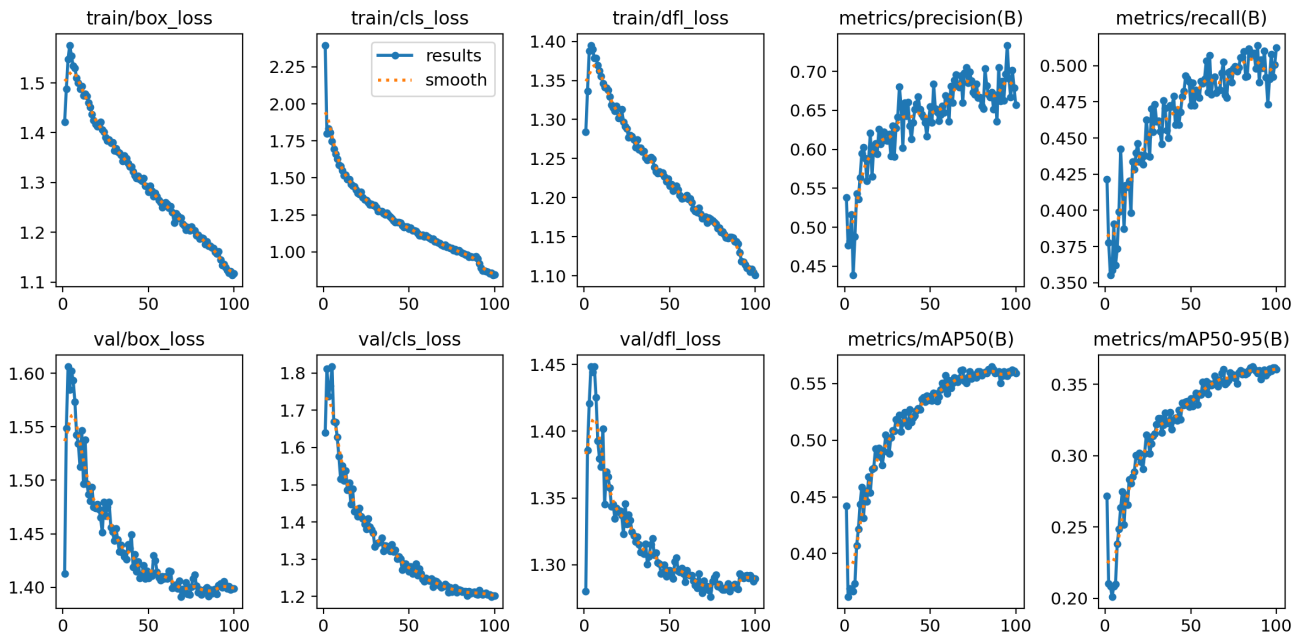


Figure 3. Our model's precision, recall, loss of box (location, detection and classification), and mAP plots.

### 4.3. Training Metrics and Performance Analysis

When training deep learning models like YOLOv8, it's crucial to understand and analyze various metrics during the training process. Key indicators such as the trend of the loss function, precision, recall, and mean Average Precision (mAP) are vital for assessing model performance (Fig. 3).

From the charts, we observe that the training set's $box\_loss$, $cls\_loss$, and $df1\_loss$ all decrease as the number of iterations increases, indicating gradual improvements in the model's ability to locate ($box\_loss$), classify ($cls\_loss$), and detect ($df1\_loss$) targets. The decline in these loss functions suggests that the model's error in recognizing and locating targets is reducing, particularly rapid at the beginning of training as the model quickly learns basic features of the data from randomly initialized weights. As training progresses, the rate of loss decline slows down as the model begins to fit more complex patterns, which requires more training iterations to achieve.

On the validation set, the trends in $box\_loss$ and $df1\_loss$ are similar to those in the training set, but $cls\_loss$ levels off after a certain point, possibly indicating that the model's learning from the training data in classification tasks is nearing saturation, or the classification tasks themselves in the data are inherently challenging, making it difficult for the model to learn further discriminative features.

Precision and recall metrics provide another perspective on model performance. The precision metric tells us how many of the targets identified by the model are correct, while recall indicates how many of the correct targets are recognized by the model. The curves for precision and recall in the charts (Fig. 5) both show an upward trend, indicating that the model's ability to recognize targets is improving over time and can detect correct targets from a more diverse set of samples.

### 4.4. Model Optimization Using F1 Score

In the field of deep learning for object detection, the F1 score is an essential metric measuring model performance. It is the harmonic mean of precision and recall, providing a comprehensive view of the model's ability to recognize positive classes (Fig. 6). The relationship graph between F1 score and the confidence threshold shows performance changes across different categories at various thresholds. This curve suggests that before a certain threshold, as the confidence threshold increases, the model becomes more accurate in identifying true targets, reducing false positives and thus increasing precision. However, setting the threshold too high makes the model overly conservative, causing many actual targets to be missed, thereby lowering the recall rate.

Different coloured curves represent different categories, and we see significant performance differences among them. For instance, the F1 score curve for the "person" category is relatively high, indicating good model performance in recognizing this category. In contrast, lower curves like that for the "train" category suggest that the model's performance in detecting such categories needs improvement.

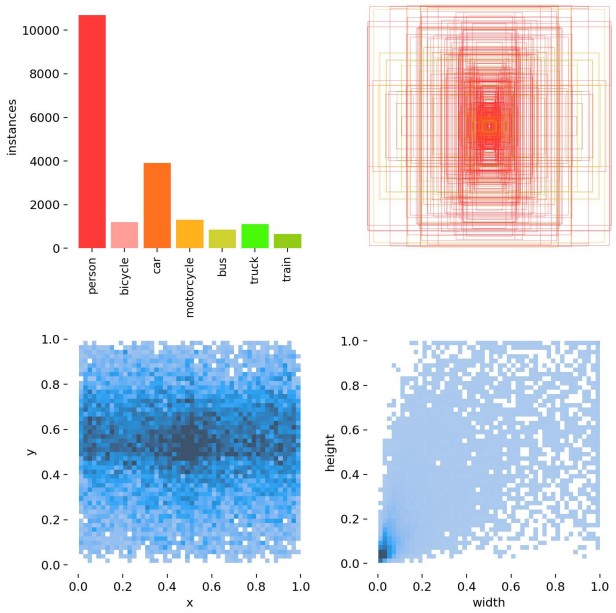


Figure 4. Dataset Concept. According to the provided dataset distribution image. The most abundant category in our dataset is humans, so the model should excel in detecting people. Additionally, the hotspots of the targets' positions in the images and the distribution of their sizes indicate that in most dataset images, the targets are concentrated in the center. The width and height distribution charts of the targets show the diversity in target sizes, which positively affects training models to recognize targets of various scales.

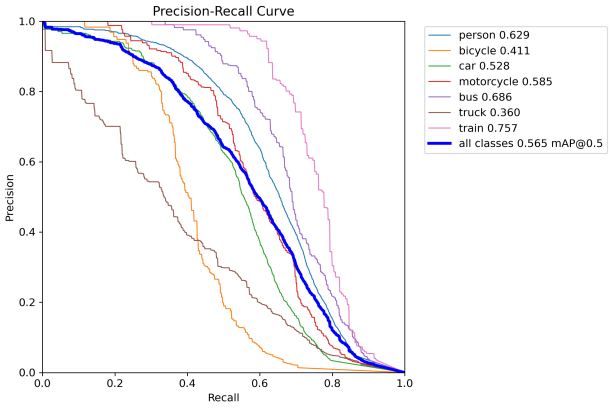


Figure 5. It is precision and recall trends. Precision measures the accuracy of positive predictions. Recall measures the ability of a model to find all the relevant cases within a dataset.

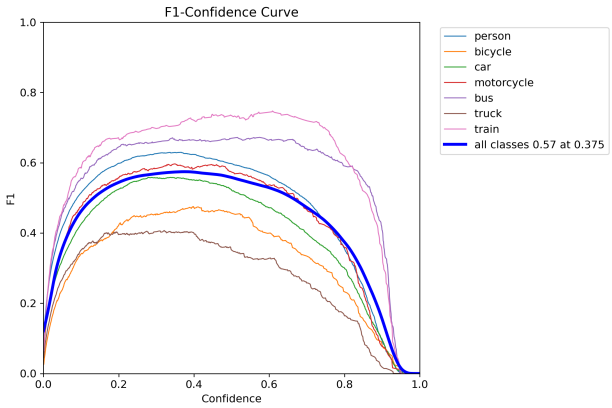


Figure 6. It is F1-confidence curve. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns by taking into account both false positives and false negatives.

Figure 7. The last frame of the video that has been selected.

Figure 8. The predicted next frame (object) based on the tracker.

This could be due to the distribution of training data, the complexity of the category itself, and the model's learning capabilities.

After utilizing the YOLO model for detection and ByteTrack for tracking, we employed the Segment Anything model's bounding box cutting to separate the target from the background layers. Subsequently, we processed the video in reverse and applied a black mask to obscure the areas outside the cut target's region. We then performed region detection to ascertain whether there were other detectable targets within the segmented target area in recent frames. If no detectable targets were present for a significant duration,

the area would be identified as standard background. This identified segment would then be excised and filled into our predicted frames to achieve the purpose of restoring the background. By utilizing the pixel/frame movement speed obtained from the previous tracker and conducting a linear regression analysis with the previous few frames (Fig. 7), we calculate the movement speed for the next frame and generate its image (Fig. 8). This process is repeated to produce a predicted video for the upcoming frames. In fact, this approach can be applied to any target object detected in a video. We use YOLO for detection and class identification, and employ a tracker (byte track) to obtain velocity data and directional trends. We implement regional monitoring to find the standard background in nearby frames, and then proceed to generate future frames for any target.

## 5. Discussion and Conclusion

We faced several obstacles in our project that were associated with object tracking distortion, such as occlusions. These elements might have an impact on our Tracking System's precision. Partial or whole occlusions would cause unexpected disturbances to object trajectories, adding complexity to continuous tracking and sometimes resulting in small errors or transient problems with object identification.

Despite the distortion-related difficulties, Our project's results hold up well; the Tracking System performs admirably in anticipating traffic and human movement and correctly detecting items like bikes, vehicles, and people. The system's robustness and efficacy in real-world traffic circumstances are highlighted by its capacity to continue operating and provide trustworthy forecasts despite distortion. This resilience attests to the stability of our methodology, validating its capacity to adjust to changing environmental circumstances and produce precise tracking and prediction outcomes, underscoring its potential for useful use in traffic management.

In practical terms, this approach significantly reduces the computational load on the tracking system during live operations, as the intensive tasks of segmenting and reconstructing frames are handled beforehand. This pre-processing not only accelerates the real-time tracking process but also ensures more efficient system operation, particularly beneficial when computational resources are limited.

Moreover, the implementation of prediction frame generation using the SAM introduces a robust tool to the field of computer vision, notably enhancing the capabilities of multi-object tracking systems. This method not only improves the accuracy of object tracking but also offers a scalable solution for managing complex scenes involving multiple interacting objects. Through this innovative approach, the system gains the ability to handle dynamic environments with enhanced precision and efficiency, showcasing a significant advancement in the practical application of computer vision technologies.

## References


[1] Gary Bishop, Greg Welch, et al. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, 8(27599-23175): 41, 2001. 4

[2] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[3] Muhammad Hussain. Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*, 11(7):677, 2023. 2

[4] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia computer science*, 199:1066–1073, 2022. 2

[5] Abdullah Ayub Khan, Asif Ali Laghari, and Shafique Ahmed Awan. Machine learning in computer vision: a review. *EAI Endorsed Transactions on Scalable Information Systems*, 8 (32):e4–e4, 2021. 1

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[7] 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. 2

[8] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[10] Christoph Sager, Christian Janiesch, and Patrick Zschech. A survey of image labelling for computer vision applications. *Journal of Business Analytics*, 4(2):91–110, 2021. 1

[11] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019. 2

[12] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond. *arXiv preprint arXiv:2305.08196*, 2023. 2

[13] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2, 3, 4