NHL STENDEN
computer vision & data science

Landmark-Based Swerving Measurement Using One-Shot Segmentation Models

NHL Stenden UAS Professorship in Computer Vision & Data Science

Alexander van Meekeren

Supervisor: Maya Aghaei Gavari

Abstract— The growth of urban cycling brings safety challenges, particularly in detecting unsafe behaviors like swerving, which can lead to collisions. Identifying these patterns in video data can improve road safety and aid infrastructure planning. This study introduces a framework for automatically measuring swerving patterns by tracking the distance between landmarks and the bike, addressing a key gap in cyclist safety research. By leveraging the Segment Anything Model (SAM) fine-tuned with PerSAM, the method involves segmenting landmarks in video frames to calculate the pixel distance between these landmarks and the bicycle, thus illustrating the swerving patterns over time. The methodology is applied to a dataset comprising 96 videos recorded in diverse cycling environments in Groningen, Netherlands, categorized into three subdatasets based on the type of bike path. The research evaluates three segmentation variants—baseline PerSAM, Multi-Class PerSAM (MCP) for 2-class segmentation, and MCP for 3-class segmentation, alongside outlier removal techniques and signal smoothing to enhance measurement accuracy. Statistical analyses, including ANOVA, reveal that the optimal configuration involves two landmarks, Isolation Forest for outlier removal, and a rolling average for signal smoothing, significantly improving measurement accuracy and robustness. Key findings indicate that while increasing the number of landmarks does not necessarily enhance accuracy due to added noise, the combination of intermediate-distance segmentation (MCP2), signal smoothing, and Isolation Forest-based outlier detection provides the most reliable and efficient approach for swerving detection. This configuration demonstrates a strong correlation with ground truth measurements, offering valuable insights for enhancing cyclist safety in real-world environments.

Index Terms— Landmark Segmentation · Signal Processing · PerSAM · Multi Class PerSAM · Bike Safety

1 Introduction

The large number of cyclists in the Netherlands has introduced significant challenges in ensuring cyclist safety [1, 2, 3, 4, 5]. One of the most pressing issues is the detection of erratic or unsafe cycling behaviors, such as swerving, which has become a critical focus of research [6, 7, 8]. Swerving, defined by sudden lateral deviations from a cyclist's intended path, can signal various underlying factors, including distraction, intoxication, fatigue, or attempts to avoid obstacles. This behavior often serves as an early indicator of more complex events, such as near-misses or collisions [9, 10, 11]. The detection of such behaviors from video data holds immense potential to improve road safety, guide infrastructure design, and support the development of advanced driver-assistance systems. This, in turn, can help better accommodate cyclists [12, 13].

The proliferation of video data from traffic cameras, dashcams, and wearable devices has opened new avenues for leveraging computer vision and machine learning to analyze cyclist swerving behavior. However, measuring swerving in video data presents unique challenges. Swerving motions are frequently subtle and may lack clear visual distinctiveness, necessitating the creation of highly sensitive and robust detection algorithms [14, 15, 16].

Accurate swerving detection depends on converting the spatial data of nearby objects, such as segmentation masks, into coordinates relative to the bicycle. This process involves measuring the distance between the cyclist and the mask of a reference landmark, like a curb, using video frame data. This transformation enables direct analysis of the rider's trajectory dynamics [17, 18]. The approach involves tracking positional shifts of the landmark over time and computing their deviation relative to the bike's position. Through this spatial-to-metric conversion, even minor swerving patterns, otherwise obscured in raw visual data, can be reliably extracted and quantified [17, 18].

This approach not only improves the precision of swerving detection but also lays the groundwork for predicting and understanding broader behavioral events in cycling scenarios [19, 20]. Advanced techniques, including optical flow analysis, trajectory prediction models, and deep learning-based landmark tracking, can facilitate this spatial-to-lateral transformation [21], enabling more accurate and reliable event detection in complex real-world environments.

However, tracking landmarks within a single reference (e.g., only a curb) can be challenging when visual obstructions occur, preventing segmentation and subsequent swerving distance calculations [22]. This research aims to automate swerving detection by leveraging multiple landmarks in video data, using spatial pixel information across several reference points to compensate for obstructions. The combined signals will be converted into lateral displacement data to accurately identify swerving patterns. This approach enables real-time event monitoring when needed. To demonstrate the feasibility of this approach, the study utilizes bicycle video data captured from a first-person perspective. This leads to the following research questions:

- How can swerving motions be measured in relation to a cyclist's surroundings?
- How can multiple swerving signals, each derived from different landmarks, be effectively combined to generate a single, cohesive swerving pattern that compensates for potential obstructions and occlusions in video frames?

It is hypothesized that distances, including the swerving of cyclists, can be measured by calculating the distance from segmented landmarks to points of interest, such as the cyclist. An increase in the number of landmarks is expected to enhance the accuracy of these

- *Alexander van Meekeren is a Computer Vision & Data Science Master student at the NHL Stenden University of Applied Sciences, E-mail: alexander.van.meekeren@student.nhlstenden.com.*
- *Maya Aghaei Gavari is a researcher at the NHL Stenden Professorship in Computer Vision & Data Science, E-mail: maya.aghaei.gavari@nhlstenden.com.*

measurements, assuming each provides independent and noise-free data. This hypothesis suggests that foundational segmentation models, with one-shot fine-tuning, are capable of effectively segmenting these landmarks, thereby demonstrating the potential of this approach to calculate distances accurately.

1.1 Related Work

Measuring swerving cycling behavior is critical for improving traffic safety and enhancing video-based surveillance systems. Early work in this domain has primarily focused on robustly detecting bicycles in low-resolution traffic videos. For instance, utilizing methods based on multi-feature and multi-frame fusion, to accurately identify bicycles even in challenging video conditions [23], thereby laying the groundwork for subsequent behaviour analysis in traffic surveillance.

Beyond mere detection, continuous research has increasingly turned to modeling the dynamics of cycling behavior to capture safety-critical actions such as swerving. Cycling swerving can be seen as a form of trajectory anomaly, a sudden deviation from the expected path of travel. In this context, recent frameworks like CyclingNet [24] have been introduced to detect cycling near misses from dashcam or bike-mounted camera videos. CyclingNet leverages spatio-temporal deep learning architectures [21] to capture abrupt trajectory anomalies, which often correspond to unsafe maneuvers such as swerving

Complementing these approaches, other methods were explored for the broader challenge of outlier detection in video data using advanced deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These methods have been applied to differentiate between normal and abnormal motion patterns in traffic scenarios, where abnormal cycling behaviors, including swerving, often indicate potential near-miss incidents [20]. While these techniques are typically designed for general outlier detection tasks, they offer valuable insights into modeling subtle variations in landmark trajectories, which are essential for accurately identifying cyclist swerving behavior.

Overall, while significant progress has been made in detecting bicycles and general anomalies in video data, specialized methods for recognizing the nuanced behavior of bicycle swerving remain an open research area. Robust swerving detection methods within video data have yet to be fully developed, particularly frameworks capable of automatically converting swerving patterns into lateral motion information for further analysis. This gap highlights the need for more advanced and automated solutions to address the challenges of measuring and interpreting cyclist swerving behavior.

1.1.1 Curb Detection

An underexplored aspect of measuring swerving behavior is linking vehicle movement to a continuous landmark within a video frame, such as the curb alongside the road. While significant research has been conducted on road detection [25, 26] and lane marking identification [27, 28], the specific relationship between swerving behavior and curb detection has received limited attention.

Traditional approaches to road and curb detection often rely on handcrafted features and geometric models. Early methods utilized edge detection, Hough transforms [29], and intensity thresholding [30, 31] to identify curbs and road boundaries. However, these techniques struggle with generalizing varying lighting conditions, occlusions, and complex urban environments. More recent approaches have shifted towards data-driven methods, particularly deep learning-based segmentation models. For instance, convolutional neural networks (CNNs) have been employed to segment road surfaces and detect curbs by learning hierarchical features from large-scale datasets [32]. Despite these advancements, curb detection remains challenging due to the subtle visual cues that differentiate curbs from surrounding surfaces [33].

Recent progress in segmentation models has opened new avenues for curb detection. The Segment Anything Model (SAM) [34], a state-of-the-art foundation model for image segmentation, has demonstrated remarkable generalization capabilities across diverse domains. SAM's ability to segment objects based on sparse user inputs or prompts makes it a promising candidate for curb detection tasks. Building on SAM, the Promptable and Efficient Segment Anything Model (PersAM) [35] further enhances segmentation efficiency and accuracy, particularly in scenarios requiring real-time processing. These models could be adapted to detect curbs by leveraging their ability to generalize across varying road conditions and perspectives.

1.1.2 Filtering Outliers in Signals

While PerSAM works effectively for single-class instances, handling multiple-class instances requires modifications to accommodate multiple reference objects. This process results in calculating distinct swerving patterns for each reference landmark relative to the bike. To merge these patterns into a cohesive output, compensating for obstructions and mitigating anomalies across multiple signals, an effective method for outlier removal is essential.

Outlier detection and removal in signal processing are critical for ensuring data reliability. Traditional approaches often rely on statistical methods (e.g., Z-score, moving average filters) [36] or machine learning models like One-Class SVM [37]. However, these methods struggle with non-stationary signals or high-frequency noise.

Wavelet transforms [38, 39] are particularly effective at capturing localized variations in both time and frequency domains, rendering them highly suitable for analyzing signals characterized by non-stationary behavior. This includes scenarios where transient phenomena (e.g., impulsive spikes or discontinuities) are present.

Another method is the Isolation Forest [40, 41], detects anomalies by exploiting their susceptibility to isolation in feature space. The algorithm constructs an ensemble of random decision trees, recursively partitioning the data by randomly selecting features and split values. Since anomalies require fewer splits to be isolated, their path lengths in the trees are shorter, allowing the algorithm to flag them as outliers.

1.2 Contribution

By addressing the lack of automated swerving detection frameworks, this study aims to advance the automation of bicycle swerving detection and provide actionable insights to enhance cyclist safety in real-world environments. This approach automates swerving pattern detection, eliminating the need for manual measurement. This is accomplished through a novel framework that integrates landmark segmentation to perform swerving calculation across multiple classes, offering a robust solution for measuring and analyzing swerving behavior.

2 Methodology

In this research, PerSAM and PerSAM extended for multiple classes are fine-tuned using reference images and corresponding landmark class masks. The fine-tuned models then segment landmarks across video frames, and the pixel distance between each landmark and the bike position is calculated. This process enables the visualization of the bike's relative distance to landmarks over time. The resulting pattern—or signal—can be interpreted as a swerving pattern, revealing insights into the bike's movement dynamics. The methodology incorporates the necessary techniques to acquire these patterns from video data.

2.1 Dataset

The dataset used in this research consists of videos recorded from the bike's point of view, with a camera mounted on the front of the bike facing the direction of movement. The dataset includes a total of 96 videos across three subdatasets: A, B, and C. All videos were recorded in the city of Groningen, Netherlands, with each subdataset representing a different cycling environment. Subdataset A contains video data from a bike lane integrated within a car road. Subdataset B features recordings from a dedicated two-way bike path, while Subdataset C consists of videos from a dedicated one-way bike path. Examples are displayed in Figure 1.



Fig. 1: Sample frames form the A, B and C dataset; dataset A features bike lanes integrated with car roads, Dataset B captures a dedicated two-way bike path, and Dataset C includes footage from a one-way bike path.

All videos contain various visual obstructions, ranging from light shadows to very heavy shadows, bus stops, and T-split intersections. Additionally, cars are sometimes parked on bike lanes, obstructing landmarks such as curbs and the bike lane itself. From time to time, the cyclist may be overtaken by other bikers or might overtake other cyclists, adding complexity to the overall visibility of landmarks in the scene.

Furthermore, all videos in the subdatasets are linked to a single bike ride. For example, videos labeled 1a, 1b, and 1c capture different scenes but share the same bike, camera, and spatial position. As a result, they also share the same calibration data for each separate bike ride.

2.1.1 Calibration Data & Annotations

Given the dataset, each bike ride has a calibration frame where calibration parameters are extracted from. The calibration data includes the following parameters for each bike ride: the bike ride index, the pixel coordinates of the bike's position, and a scale factor used to convert pixel distances to real-world lateral distances. The calibration values are provided in Appendix A Table 2.

For each bike ride, a calibration frame is used; an example is shown in Figure 2. A calibration stick helps determine the calibration values. Since the bike's position in the annotations corresponds to the middle of the stick, these pixel coordinates are used to compute the swerving distance.



Fig. 2: Calibration frame for bike ride 1.

The annotations for every video, spanning all scenes, include partially annotated segments. All the annotations are created by traffic safety domain experts. For Scene A, annotations are available for 22.6% across all the frames from 18 out of 31 videos. Similarly, for Scene B, annotations cover 33.7% across all the frames from 18 out of 32 videos, and for Scene C, 24.3% across all the frames from 18 out of 32 videos. An overview is provided in Table 1.

Scene	Frames Annotated	Videos Annotated
A	9593 (22.6%)	18/31
B	10491 (33.7%)	18/32
C	11493 (24.3%)	18/31

Table 1: Overview of annotation metadata across different datasets.

The annotations, which were carefully labeled by hand across 30,000 frames, include the lateral distance between the bike and the curb within each video frame. Specifically, these annotations were obtained by traffic researchers at Rijksuniversiteit Groningen. They measured the pixel distance between the bike and the curb in the video frames. By using a calibration stick placed in front of the bike in each video, where the size of the stick is known, the pixel distance was converted into the actual lateral distance.

2.2 Landmark Segmentation

To segment curbs and determine the swerving pattern, this research employs *Segment Anything* (SAM) [34], a foundation model for image segmentation that generalizes to new objects and scenes without task-specific training. SAM comprises three key components: an image encoder E, a prompt encoder P, and a mask decoder D. These components work together to generate segmentation masks based on an input image and a prompt.

Given an input image I, the image encoder extracts high-dimensional features, represented as:

$$F = E(I), \tag{1}$$

where $F \in \mathbb{R}^{h \times w \times d}$ denotes the feature map with dimensions h, w, and feature depth d. The prompt encoder P processes user-provided prompts p, such as points, boxes, or text, to produce an embedding representation:

$$P(p) = \mathbf{z}_p, \tag{2}$$

where $\mathbf{z}_p$ is the prompt embedding. The mask decoder D then generates a segmentation mask M using the feature map and prompt

embedding:

$$M = D(F, \mathbf{z}_p). \tag{3}$$

PerSAM (Personalized Segment Anything Model) [35] extends SAM by enabling one-shot personalization, allowing automatic segmentation of a user-specified landmark in new images or videos without repeated prompting. Given just a single reference image and a mask (either precise or rough), PersAM adapts SAM to recognize the target object by leveraging feature matching and similarity learning [35].

To further enhance segmentation accuracy across varying object scales, by default, SAM selects the most appropriate scale for the given prompt, where scale refers to the relative size or resolution at which objects are detected in an image, influencing how large or small features are perceived. However, the multi-scale outputs can also be leveraged for refined analysis, such as fusing predictions to enhance segmentation accuracy across varying object sizes. PersAM introduces PerSAM-Finetune [35], a scale-aware fine-tuning variant that refines mask predictions through a weighted fusion of three-scale masks (M_1, M_2, M_3) generated by SAM. The final mask M is computed as in Equation 4

$$M = w_1 \cdot M_1 + w_2 \cdot M_2 + (1 - w_1 - w_2) \cdot M_3 \tag{4}$$

Where w_1 and w_2 are learnable weights initialized at $1/3$ and fine-tuned on the reference mask. While keeping SAM's pretrained weights frozen to preserve its generalization, this lightweight adaptation allows PerSAM-Finetune to dynamically select the optimal segmentation scale for different objects.

This research presents a multi-class generalization of PerSAM (Multi-Class PerSAM, MCP) through class-specific parameterization of all adaptive components, including both feature similarity computation and multi-scale mask fusion. Here, a class-conditioned reformulation of the mask aggregation process (Equation 5), where distinct scale-weighting parameters are learned for each target class $x \in \{1, \ldots, C\}$.

$$M^{(x)} = w_1^{(x)} \cdot M_1^{(x)} + w_2^{(x)} \cdot M_2^{(x)} + (1 - w_1^{(x)} - w_2^{(x)}) \cdot M_3^{(x)} \tag{5}$$

2.3 Swerving Measuring Process

The training process for Multi-Class PerSAM and PerSAM involves processing each video frame by predicting landmarks and extracting their pixel positions. Outlier removal is performed before scaling the landmark curb signal to match available annotations representing distances from the curb, using a calculated scaling factor. For Multi-Class PerSAM, all signals are first scaled to the landmark curb signal, then combined through mean aggregation. A rolling average is applied to smooth the resulting signal, producing a clear swerving pattern that captures the vehicle's motion trajectory. This process can be visualized in Figure 3. Each stage will be examined in detail in this subsection, beginning with the model training procedures.

2.3.1 (Multi Class) PerSAM Training

The training procedure followed the official PerSAM implementation [35], with both single-class and multi-class variants trained for 1000 epochs [35]. The standard PerSAM model ($P(1)$) utilized the right curb in each scene as the training image and reference landmark, selected for consistent annotation availability across all scenes to enable direct comparison.

Two multi-class configurations were implemented: a 2-class variant ($MCP(2)$) and a 3-class variant ($MCP(3)$). Additional landmarks were progressively selected at increasing distances from the primary curb. Both multi-class models incorporated the main bike lane as an extra class, chosen for each scene based on its proximity to the primary curb, potentially providing supplementary segmentation cues when curb detection proves challenging.

The MCP3 variant included a third landmark representing each scene's leftmost boundary: the adjacent main road in Scene A, a left divider separating the bike lane from the main road in Scene B, and an additional left curb bordering greenery in Scene C. An example of an segmentation attempt can be seen in Figure 4, where three landmark masks were predicted.

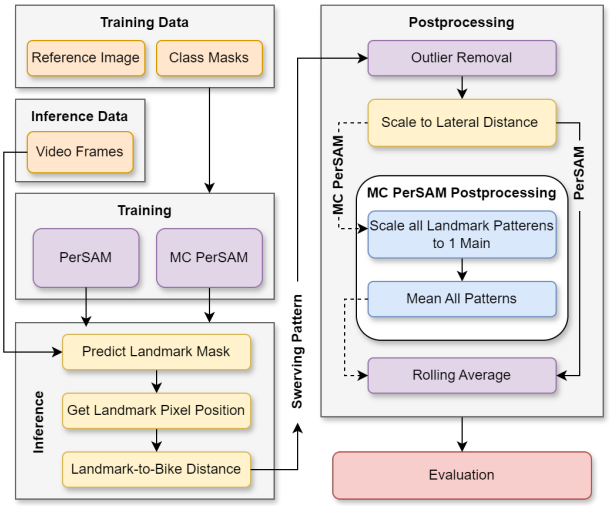

Fig. 3: Schematic representation of the swerving measuring process.

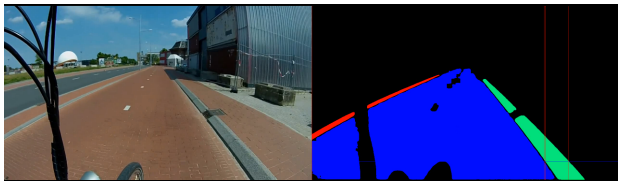

Fig. 4: Example image of the segmentation of multiple landmarks within a frame using Multi Class PerSAM. Left: example frame, Right: predicted landmark masks

All model variants were trained for each scene with their weights preserved for subsequent experimentation.

This research focuses on the application rather than the architectural improvement of PerSAM. Consequently, no segmentation optimization attempts were made, nor were segmentation results evaluated. Referring to the original PerSAM paper, the mean Intersection over Union (mIoU), when evaluated on the PerSEG dataset [35], showed that fine-tuning PerSAM for one cone class resulted in an mIoU of 95.3 [35], indicating the effectiveness of this approach. Potential performance enhancements would require either PerSAM architectural modifications or SAM backbone fine-tuning, both beyond the study's scope.

2.3.2 Landmark Position Extraction

Following the training of segmentation models, all of the frames from the dataset videos were processed using all three model variants - $P1, MCP2, and MCP3$ - to predict landmark locations by generating corresponding segmentation masks. The landmark extraction process utilizes the calibration scanline height y-position to determine pixel coordinates, identifying the pixel closest to the curb at the specified calibration height. This measured pixel position is then compared against the (x, y) calibration coordinates, with the resulting difference representing the pixel distance D between the landmark and the bicycle's position. This process is outlined in the algorithm below. After processing all video frames and computing the distance measurements D, the output constitutes raw swerving signals.

Require: Video frames F, calibration coordinates y_{cal}, x_{cal}
Ensure: Distance matrix D containing all measurements

for each frame $f_i \in F$ **do**
  for each model $m_j \in \{P1, MCP2, MCP3\}$ **do**
    $M_{i,j} \leftarrow \text{GenerateSegmentationMask}(m_j, f_i)$
    for each landmark l_k in $M_{i,j}$ **do**
      $P \leftarrow \text{GetPixelsAtY}(l_k, y_{cal})$
      $p_{closest} \leftarrow \underset{p \in P}{\text{argmin}}(\text{DistanceToCurb}(p))$
      $D_{i,j,k} \leftarrow |p_{closest}.x - x_{cal}|$
    end for
  end for
end for
return D

2.3.3 Outlier Removal

The acquired raw swerving signals undergo an outlier removal process where different approaches are tested in this research. The first approach is an isolation forest algorithm [40, 41] that identifies outliers through recursive partitioning of the feature space, where the anomaly score derives from the expected path length, shown in Equation 6

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{6}$$

The function $s(x,n)$ represents the anomaly score for a data point x given a sample size of n. The term $E(h(x))$ denotes the expected path length for point x across all isolation trees in the forest. The function $c(n)$ refers to the average path length of an unsuccessful search in a Binary Search Tree (BST) containing n nodes.

The second approach employs Daubechies wavelets [42], a widely used family of wavelets known for their compact support, orthogonality, and vanishing moments. These properties enable precise time-frequency localization, making Daubechies wavelets particularly effective for tasks such as noise reduction, outlier removal, and multi-scale signal analysis.

Both outlier removal methods employed a 5% threshold criterion. This choice was motivated by the following key considerations: First, under the assumption of Gaussian-distributed noise, approximately 5% of the wavelet coefficients are expected to exceed 2 standard deviations from the mean. Second, this threshold empirically maintains 95% of legitimate signal components while effectively eliminating significant artifacts.

2.3.4 Signal Scaling and Alignment

The research utilizes annotations spanning from the bicycle position to the curb in each frame. All landmark detections corresponding to the same curb are converted from pixel distances to real-world lateral distances through multiplication by a scaling factor. While this factor could be calculated per-video using dedicated calibration values, this implementation automatically derives it from the first annotated frame under the assumption of accurate initial predictions, an assumption verified through manual inspection. The scaling factor is computed as follows in Equation 7

$$scalingfactor = \frac{real - worlddistance(cm)}{pixeldistance(px)} \tag{7}$$

The multi-class PerSAM outputs generate multiple swerving signals, including both primary curb detections and secondary landmarks like bikelanes. To combine these signals effectively, the system first computes the mean lateral offset between each secondary signal and the primary curb reference. Each secondary signal is then adjusted by subtracting its characteristic mean offset, bringing all trajectories into consistent alignment with the curb-based coordinate system. Finally, the algorithm averages all aligned signals to produce a single, unified swerving trajectory that incorporates information from all detectable landmarks.

2.3.5 Signal Smoothing

A final smoothing stage is applied to the measured swerving signal using a rolling average with a window size of 5 frames. A smaller window size might not sufficiently smooth the signal, while a larger one could overly flatten it, eliminating subtle swerving deviations. This effect can be observed in Figure 5. This temporal filtering reduces high-frequency noise while preserving the underlying swerving patterns, resulting in smoother trajectory measurements that better reflect the cyclist's actual lateral movements.

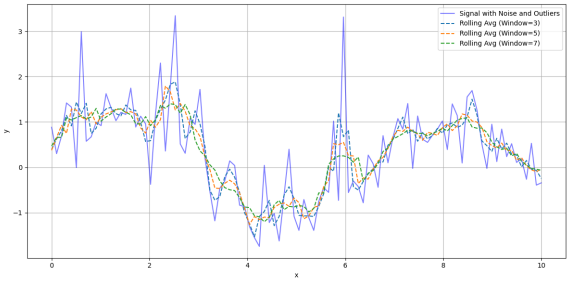

Fig. 5: A signal (solid blue), which includes outliers and noise, is smoothed using rolling averages with varying window sizes; 3 (dotted blue), 5 (dotted orange), and 7 (dotted green).

2.4 Evaluation

To evaluate the performance of the predicted bike swerving trajectory against the annotated ground truth, the following metrics were employed, each providing unique insights into different aspects of the prediction accuracy:

The Root Mean Square Error (RMSE), Equation 8 [43], measures the standard deviation of prediction errors, giving more weight to large errors due to the squaring operation. This is particularly valuable for swerving measurements as it penalizes significant deviations more heavily. Here, y_i is the annotated swerving distance, $\hat{y}_i$ is the predicted swerving distance, and n is the number of samples.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

The Mean Absolute Error (MAE), Equation 9 [44], provides a linear score of average prediction error magnitude, being more robust to outliers than RMSE. This metric gives a straightforward understanding of the typical prediction error in physical units (e.g., lateral distance), which helps assess the practical significance of errors in swerving distance prediction.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

The Maximum Absolute Error, Equation 10, identifies the single worst prediction error in a given prediction signal, highlighting local performance extremes.

$$\text{MaxAE} = \max(|y_i - \hat{y}_i|) \tag{10}$$

Spearman's ρ, Equation 11 [45], assesses the relationship between the predicted and the actual swerving movements rank in the same order, where d_i is the difference between ranks of corresponding values. Unlike other metrics, it does not require the relationship to be perfectly linear, it checks if bigger swerves in the annotation generally correspond to bigger swerves in the prediction. Without assuming a linear relationship, Spearman's ρ evaluates whether swerving events in the predictions consistently align with the annotated ground truth, assessing rank-order agreement rather than strict numerical precision.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{11}$$

2.4.1 ANOVA

Following the computation of evaluation metrics across all three datasets, an analysis of variance (ANOVA) was conducted to assess the statistical significance of each experimental factor on system performance. This rigorous statistical examination evaluated the impact of three key independent variables: (1) the segmentation methodology (MCP2, MCP3, or P1), (2) the application of rolling average smoothing, and (3) the choice of outlier removal technique (Isolation Forest, Wavelet, or none). The ANOVA considered all four performance metrics (MAE, RMSE, Max AE, and Spearman's ρ) as dependent variables, enabling a comprehensive assessment of each factor's influence across multiple dimensions of system performance.

3 Experiments, Results & Discussion

Experiments in this research involved three segmentation variants for bicycle swerving analysis: the baseline PerSAM model, Multi-Class PerSAM (MCP) configured for 2-class segmentation, and MCP configured for 3-class segmentation. Training procedures involved one-frame processing of three distinct video datasets representing diverse cycling scenarios.

The segmentation landmark outputs functioned as reference points for lateral distance computation in the swerving detection pipeline. Generated masks enabled frame-by-frame displacement measurements between a bicycle and detected landmarks, with temporal aggregation of these measurements quantifying swerving magnitude throughout video sequences. Optionally, outlier removal techniques are employed on the measured swerving signal to further enhance accuracy.

3.1 Experimental Setup

To systematically evaluate the performance of the proposed swerving detection framework, this research utilized a full factorial experimental design examining three critical factors. First, the landmark reference segmentation method was varied between MCP (Multi-Class PerSAM) with 2 and 3 classes, and P (PerSAM) with only one class, with each configuration influencing the bike-to-landmark distance estimation in distinct ways. Second, the effect of signal smoothing through application of a rolling average with a 5-frame window size is tested to assess its impact on noise reduction versus real-time responsiveness. Third, three outlier detection approaches were compared: Isolation Forest with a contamination parameter of 0.05 for model-based outlier detection, a Wavelet Transform with a threshold of 0.05 for frequency-domain filtering, and a baseline configuration without any outlier filtering.

The complete experimental design, comprising all 18 possible combinations of these factors, is detailed in Appendix B.1 Table 3. Each configuration was evaluated across three distinct datasets (A, B, and C) representing diverse urban cycling scenarios that varied in obstacle density, lighting conditions, and traffic patterns. Performance was assessed using four quantitative metrics: MAE, measuring the average deviation in centimeters from ground truth annotations. RMSE, emphasizing larger deviations. Max AE, identifying worst-case detection failures, and Spearman's rank correlation coefficient, evaluating trend agreement between detected and actual swerving patterns. Ground truth data was manually annotated by domain experts, ensuring reliable benchmarking.

3.2 Quantitative results

The proposed experiments were evaluated across three distinct datasets (A, B, and C) under varying experimental configurations, with performance metrics summarized in Tables 5, 6, and 7. An aggregated comparison of all methods and datasets is presented in Table 4, providing a comprehensive assessment of swerving measurement performance under different parameterizations.

Across all datasets, the MCP2 method with rolling average enabled and Isolation Forest outlier detection consistently achieved the lowest errors, with a mean absolute error (MAE) of 2.68 cm, root mean squared error (RMSE) of 3.92 cm, and a maximum absolute error (Max AE) of 19.43 cm. This configuration also exhibited the highest Spearman's rank correlation coefficient (ρ = 0.93), indicating strong correlation between predicted and observed values. The inclusion of rolling averaging generally improved performance, reducing MAE compared to non-rolling configurations. Isolation Forest outlier detection outperformed both wavelet-based and no outlier detection approaches, reducing MAE compared to the no-detection baseline.

Dataset-specific analyses revealed notable variations in performance. Dataset B exhibited the lowest errors overall, with MCP2 achieving an MAE of 2.16 cm when using rolling averages and Isolation Forest. In contrast, Dataset C proved more challenging, with higher errors across all configurations, particularly for the P1method without outlier detection (MAE = 9.67 cm, RMSE = 24.89 cm). The MCP3 method demonstrated intermediate performance, with MAE values ranging from 3.72 to 6.24 cm, suggesting that increasing the model complexity beyond 2 classes did not yield consistent improvements.

The wavelet-based outlier detection method showed mixed results, occasionally improving upon no outlier detection but generally underperforming compared to Isolation Forest. For instance, in Dataset A, wavelet denoising reduced MAE by 7.6% compared to no outlier detection but remained 15.1% higher than Isolation Forest. The maximum error metrics further highlighted the robustness of Isolation Forest, which consistently mitigated extreme prediction errors.

These findings suggest that the combination of MCP2 with rolling averaging and Isolation Forest outlier detection provides the most reliable and accurate predictions across diverse datasets. The consistent performance gains from rolling averaging indicate its utility in smoothing temporal fluctuations, while the superior outlier rejection capability of Isolation Forest underscores its importance in handling noisy real-world data.

3.2.1 Statistical Analysis

The statistical analysis reveals significant effects across all experimental factors for each performance metric, with particularly strong influences from segmentation methods and outlier removal techniques. The results demonstrate metric-specific patterns that inform optimal system configuration.

For Mean Absolute Error, method selection shows highly significant effects ($p<0.001$), with both alternative methods increasing error relative to the baseline MCP2. The P method shows the strongest degradation (β=1.708), followed by MCP3 (β=0.941). Isolation Forest emerges as the most effective outlier method, reducing MAE by 1.698 cm ($p<0.001$), while wavelet filtering provides a smaller but still significant improvement (β=-0.718). Rolling average shows no significant MAE effect (p=0.106).

The RMSE analysis reveals even more pronounced method effects (p=0.002), with the P method increasing error by 4.792 cm ($p<0.001$). Notably, rolling average demonstrates significant RMSE reduction (β=-2.123), suggesting it improves error consistency despite its neutral MAE effect. Isolation Forest again shows strong benefits (β=-4.808, $p<0.001$), while wavelet filtering lacks significance (p=0.183).

Maximum Absolute Error results show dramatic impacts from both processing techniques. Rolling average produces remarkable reductions (β=-27.147), as does Isolation Forest (β=-45.252). Method effects remain significant (p=0.011), with the P method increasing maximum error by 35.695 cm (p=0.003).

For Spearman's ρ, method choice shows extremely strong effects ($p<0.001$), with MCP3 particularly detrimental (β=-0.051). Both outlier methods significantly improve correlation, with Isolation Forest showing greater benefit (β=0.027) than wavelet (β=0.017). Rolling average shows borderline significance (p=0.056), suggesting a mild positive trend.

The analysis reveals several key insights about system performance. Most notably, the baseline MCP2 method demonstrates consistent superiority across all evaluated metrics compared to alternative approaches. Isolation Forest emerges as particularly effective, showing robust improvements for every error measure and

correlation coefficient. Rolling average processing proves especially valuable for two specific aspects: substantially reducing maximum absolute errors and enhancing measurement consistency as reflected in RMSE scores. While wavelet filtering shows more limited effects, it still provides statistically significant, although minimal, improvements in correlation measurements. The influence of methodological choices appears strongest when examining correlation coefficients (ρ) and mean absolute error values.

Based on these findings, the optimal system configuration combines MCP2 with Isolation Forest outlier removal and rolling average processing. This combination delivers balanced performance enhancements across all measurement dimensions. The conclusions are particularly robust given the highly significant p-values ($p<0.01$ for most critical effects) and the consistent patterns observed across different performance metrics. This configuration therefore represents the most reliable choice for achieving comprehensive system optimization. The complete statistical results, presented in Appendix B.3 Tables 8-11, provide detailed coefficients and confidence intervals for all factor combinations. Residual analysis confirmed model assumptions were satisfied across all tests, supporting the validity of these findings.

3.3 Qualitative Analysis

Building upon the static and quantitative analyses, the results are further supported by qualitative findings. The enhanced performance of MCP2 over P1 is exemplified in a video from scene C, as illustrated in Figure 6. Utilizing the optimal configuration of signal post-processing, which includes a rolling average and an isolation forest, it is evident that the right graph depicting the swerving pattern of MCP2 exhibits reduced noise overall compared to when only one class is used. This highlights the potential of using more than one reference landmark for serving detection.

Additionally, in the same figure, between frames 400 and 500, the reference landmark, the curb, is occasionally obstructed when using PerSAM, resulting in gaps in the prediction. However, with MCP2, other signals are available for the prediction to rely on, leading to a complete swerving prediction without gaps.

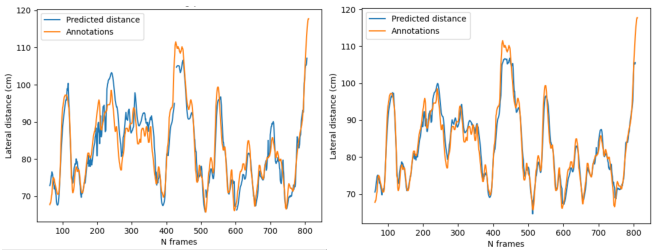

Fig. 6: Swerving plot example. Blue represents measured swerving distances, and orange the annotations. Left: Swerving measurements given PerSAM trained on only one class. Right: Acquired swerving measurements using Multi-Class PerSAM with 2 classes.

However, establishing that MCP2 has better accuracy than P1 does not imply that MCP2 always results in high accuracy. Occasionally, as seen in Figure 7 (right), the swerving measurements of MCP2, although better than those of P1 (left), are not highly accurate.

Upon examining the raw segmentation masks of P1, as shown in Figure 8, it is evident that there are heavy shadows in the curb segmentation area due to the unfortunate placement of the biker's shadow. This results in mask predictions with a high error rate.

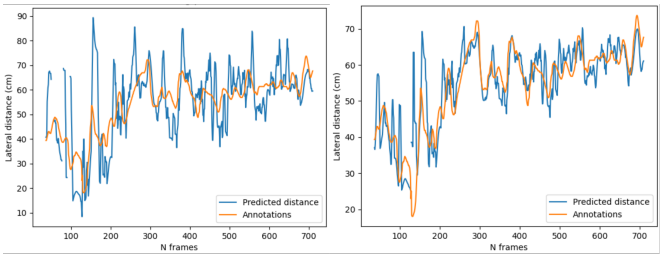

Fig. 7: Swerving plot example. Blue represents measured swerving distances, and orange represents annotations. Left: Swerving measurements given PerSAM trained on only one class. Right: Acquired swerving measurements using Multi-Class PerSAM 2 with two classes.

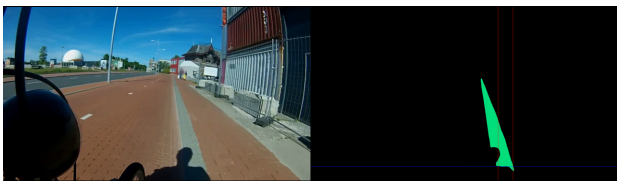

Fig. 8: Segmentation masks of P1 showing heavy shadows affecting the curb segmentation area.

Similarly, in Figure 9, the same issue is observed with MCP2, where the heavy shadow of the biker overlaps both the segmentation areas of the curb and the bike lane. Even the addition of an extra class, thus using MCP3, is of no help in this example, as the signal and image are blocked by a bike lamp at the location where the third landmark would be.

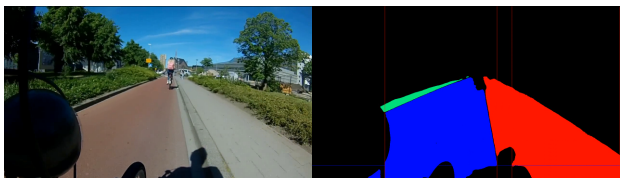

Fig. 9: Segmentation masks of MCP2 showing heavy shadows affecting both the curb and bike lane segmentation areas.

All of this indicates the importance of choosing the right reference landmarks when conducting segmentation and measurement tasks. Selecting landmarks that are less prone to obstructions and shadows can significantly improve the accuracy of the measurements. While MCP2 offers improvements over P1, attention to setup details and the choice of reference landmarks remain key factors in maximizing the effectiveness of the system.

4 Conclusion

In this study, the SAM model was fine-tuned using PerSAM with reference images and landmark class masks to segment landmarks in video frames. By measuring the distance between landmarks and the bike, the bike's swerving pattern was visualized, offering insights into its movement. The methodology includes techniques to extract these patterns from video data. Experiments used three segmentation variants for bicycle swerving analysis: the baseline PerSAM model, Multi-Class PerSAM (MCP) for 2-class segmentation, and MCP for 3-class segmentation. In addition, outlier removal techniques and signal smoothing are applied to reduce error, resulting in quantified swerving magnitude throughout the videos.

With this process completed, we can address the research questions: **"How can swerving motions be measured in relation to a cyclist's surroundings?"**

By segmenting video landmarks per frame and calculating the pixel distance from the curb to the bike, we can detect swerving patterns. Converting pixel distance to lateral distance using calibration parameters and aggregating all landmark distances yields a swerving pattern of the cyclists.

Statistical analysis suggests that two landmarks are optimal for measuring swerving motion. This contradicts the hypothesis that more landmarks improve measurement accuracy by reducing error. Additional landmarks can introduce noise, and adding a third landmark degrades signal performance. The effect of different segmentation methods can also been seen in statistical analyses across datasets. For MAE, the P method shows the most significant degradation (1.708 cm) compared to MCP2, followed by MCP3 (0.941 cm). The Maximum Absolute Error is also notably affected, with the P method increasing the maximum error by 35.695 cm. Spearman's ρ indicates strong method effects, with MCP3 being particularly detrimental.

"How can multiple swerving signals, each derived from different landmarks, be effectively combined to generate a single, cohesive swerving pattern that compensates for potential obstructions and occlusions in video frames?"

To optimally combine multiple swerving signals, start by using Isolation Forest with a 5% outlier rate to remove outliers. Scale all signals to a chosen reference signal, such as the curb landmark swerving signal, and average them. Apply a rolling average with a window of 5 to smooth the signals.

This method of outlier removal and rolling average provides the most representative temporal distance measurements, supported by statistical analyses. Isolation Forest reduces the MAE by 1.698 cm, while the wavelet-based method improves it by 0.718 cm. The rolling average does not significantly affect MAE. For Maximum Absolute Error, the rolling average reduces error by 27.147 cm, and Isolation Forest by 45.252 cm. For Spearman's ρ, Isolation Forest improves correlation, with wavelet filtering showing a smaller benefit, and the rolling average showing borderline significance.

Overall, while absolute performance varies across datasets, the relative benefits of specific configurations remain consistent. Combining intermediate-distance segmentation (MCP2), signal smoothing, and Isolation Forest-based outlier detection proves to be the most reliable approach. This configuration balances accuracy, robustness, and computational efficiency across diverse cycling scenarios and shows a strong correlation with ground truth measurements (mean $\rho = 0.93$), making it particularly useful for safety-critical applications requiring reliable swerving detection.

5 Future Work

While this study offers valuable insights into distance measurement through video segmentation and signal processing techniques, several promising directions remain for future research.

First, conducting an additional analysis of supplementary landmarks could help identify specific conditions or environments where these landmarks enhance swerving measurement accuracy without adding noise. Primary landmarks—such as lane markings or curbs-can sometimes be obstructed by parked vehicles, pedestrians, or debris. In such cases, dynamically selecting landmarks based on real-time visibility, environmental context, or cyclist behavior could improve system robustness. This adaptive approach may lead to more precise and reliable measurements, especially in complex urban environments where static landmarks are not consistently available.

Second, advanced signal processing techniques could further improve robustness. Due to time constraints, it was not feasible to implement these enhancements. However, developing advanced outlier detection and smoothing algorithms, or utilizing machine learning for adaptive real-time parameter optimization, could potentially refine the analysis of swerving patterns.

Finally, broader application and validation of the proposed methodologies across a wider range of scenarios—such as urban bike lanes, high-traffic roads, and shared pedestrian-cyclist spaces—would enhance their generalizability and practical relevance. The current analysis is constrained to three scenes within a single urban setting, which may not reflect the full variability of real-world conditions, including differences in infrastructure, traffic density, and cyclist behavior across cities. Expanding the scope of validation would build on the foundation established here, contributing to both theoretical advancements and improved real-world cyclist safety outcomes.

6 Broader Impact

The broader impact of measuring the swerving behavior of cyclists through video data encompasses significant ethical and social dimensions. Ethically, the use of segmentation algorithms to measure swerving behavior raises critical concerns. One major issue is the potential for these algorithms to be used for malicious tracking. Without proper safeguards, the technology could be exploited to monitor individuals without their consent, leading to serious privacy violations. To mitigate this risk, it is essential to implement stringent regulations and oversight mechanisms to ensure that the technology is used solely for its intended purpose.

In regulated monitoring environments, safeguarding individual privacy must be a top priority. Those being monitored—such as cyclists, should be clearly informed about what data is being collected and how it will be used. At the same time, the system may inadvertently capture information about third parties, such as bystanders, vehicle license plates, or other identifiable details. To address these concerns, strong anonymization methods should be implemented to protect all individuals' identities. Additionally, data must be securely stored to prevent unauthorized access. Ensuring transparency in the monitoring process and obtaining informed consent from primary participants are essential components of ethical data collection.

Socially, this research can significantly enhance road safety by identifying high-risk areas and behaviors, leading to better infrastructure and targeted safety measures. By analyzing swerving patterns, city planners and traffic engineers can design safer bike lanes and intersections, reducing the likelihood of accidents. Additionally, the data can be used to educate cyclists about safe riding practices, fostering a culture of safety and responsibility. Public awareness campaigns based on the research findings can encourage cyclists to adopt safer behaviors, ultimately contributing to a safer cycling environment for everyone.

While the measurement of cyclist swerving through video data offers substantial benefits for road safety, it is crucial to address the ethical implications carefully. By implementing robust regulations, ensuring privacy and consent, and maintaining transparency, the broader impact of this technology can be maximized while minimizing potential risks.

Acknowledgements

This project is financially supported by Regieorgaan SIA (part of NWO) under the RAAK Publiek project BikeSafeAI.

REFERENCES

[1] Paul Schepers, Divera Twisk, Elliot Fishman, Aslak Fyhri, and Anne Jensen. The dutch road to a high level of cycling safety. *Safety science*, 92:264–273, 2017.

[2] Paul Schepers, Marco Helbich, Marjan Hagenzieker, Bas de Geus, Marco Dozza, Niels Agerholm, Anna Niska, Noora Airaksinen, Francis Papon, Regine Gerike, et al. The development of cycling in european countries since 1990. *European Journal of Transport and Infrastructure Research*, 21(2):41–70, 2021.

[3] Sonja Haustein and Mette Møller. E-bike safety: Individual-level factors and incident characteristics. *Journal of Transport & Health*, 3(3):386–394, 2016.

[4] Michael Branion-Calles, Trisalyn Nelson, and Meghan Winters. Comparing crowdsourced near-miss and collision cycling data and official bike safety reporting. *Transportation research record*, 2662(1):1–11, 2017.

[5] Mohamed Bayoumi Kamel and Tarek Sayed. The impact of bike network indicators on bike kilometers traveled and bike safety: A network theory approach. *Environment and Planning B: Urban Analytics and City Science*, 48(7):2055–2072, 2021.

[6] Masayuki Miyadai, Teruo Uetake, and Masahiro Shimoda. How does a cyclist avoid obstacles? *Journal of human ergology*, 41(1_2):95–100, 2012.

[7] ZHAO Yuqing. Analysis of driver and cyclist responses in car-to-cyclist conflicts.

[8] Frank Westerhuis. Advancing the age of cycling. 2021.

[9] Zaili Yang, Zhisen Yang, John Smith, and Bostock Adam Peter Robert. Risk analysis of bicycle accidents: A bayesian approach. *Reliability Engineering & System Safety*, 209:107460, 2021.

[10] Paul M Salmon, Mitch Naughton, Adam Hulme, and Scott McLean. Bicycle crash contributory factors: A systematic review. *Safety science*, 145:105511, 2022.

[11] Antonella Scarano, Massimo Aria, Filomena Mauriello, Maria Rella Riccardi, and Alfonso Montella. Systematic literature review of 10 years of cyclist safety research. *Accident Analysis & Prevention*, 184:106996, 2023.

[12] Adnan Shaout, Dominic Colella, and Selim Awad. Advanced driver assistance systems-past, present and future. In *2011 Seventh International Computer Engineering Conference (ICENCO'2011)*, pages 72–82. IEEE, 2011.

[13] Vipin Kumar Kukkala, Jordan Tunnell, Sudeep Pasricha, and Thomas Bradley. Advanced driver-assistance systems: A path toward autonomous vehicles. *IEEE Consumer Electronics Magazine*, 7(5):18–25, 2018.

[14] Frank Westerhuis and Dick De Waard. Using commercial gps action cameras for gathering naturalistic cycling data. *Journal of the Society of Instrument and Control Engineers*, 55(5):422–430, 2016.

[15] Ryan Toxopeus, Shady Attalla, Sam Kodsi, and Michele Oliver. Driver response time to cyclist path intrusions. Technical report, SAE Technical Paper, 2018.

[16] Yan Ke. *A Cyclist Detection and Tracking System for Heavy Goods Vehicles*. PhD thesis, 2022.

[17] Ryan De Iaco, Stephen L Smith, and Krzysztof Czarnecki. Safe swerve maneuvers for autonomous driving. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1941–1948. IEEE, 2020.

[18] Ryan De Iaco, Stephen L Smith, and Krzysztof Czarnecki. Universally safe swerve maneuvers for autonomous driving. *IEEE Open Journal of Intelligent Transportation Systems*, 2:482–494, 2021.

[19] Darren Boss, Trisalyn Nelson, and Meghan Winters. Monitoring city wide patterns of cycling safety. *Accident Analysis & Prevention*, 111:101–108, 2018.

[20] Shumayla Yaqoob, Salvatore Cafiso, Giacomo Morabito, and Giuseppina Pappalardo. Detection of anomalies in cycling behavior with convolutional neural network and deep learning. *European transport research review*, 15(1):9, 2023.

[21] Matthew J Williams and Mirco Musolesi. Spatio-temporal networks: reachability, centrality and robustness. *Royal Society open science*, 3(6):160196, 2016.

[22] Yixin Zhang, Nicholas Konz, Kevin Kramer, and Maciej A. Mazurowski. Quantifying the limits of segmentation foundation models: Modeling challenges in segmenting tree-like and low-contrast objects, 2025.

[23] Yicheng Zhang and Qiang Ling. Bicycle detection based on multi-feature and multi-frame fusion in low-resolution traffic videos, 2017.

[24] Mohamed R Ibrahim, James Haworth, Nicola Christie, and Tao Cheng. Cyclingnet: Detecting cycling near misses from video streams in complex urban scenes with deep learning. *IET Intelligent Transport Systems*, 15(10):1331–1344, 2021.

[25] Hui Kong, Jean-Yves Audibert, and Jean Ponce. General road detection from a single image. *IEEE Transactions on Image Processing*, 19(8):2211–2220, 2010.

[26] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE, 2013.

[27] Youcheng Zhang, Zongqing Lu, Xuechen Zhang, Jing-Hao Xue, and Qingmin Liao. Deep learning in lane marking detection: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):5976–5992, 2021.

[28] Mauricio Braga De Paula and Claudio Rosito Jung. Real-time detection and classification of road lane markings. In *2013 XXVI conference on graphics, patterns and images*, pages 83–90. IEEE, 2013.

[29] Weifeng Liu, Zhenqing Zhang, Shuying Li, and Dapeng Tao. Road detection by using a generalized hough transform. *Remote Sensing*, 9(6):590, 2017.

[30] José M Álvarez Alvarez and Antonio M Ĺopez. Road detection based on illuminant invariance. *IEEE transactions on intelligent transportation systems*, 12(1):184–193, 2010.

[31] Alberto Hata and Denis Wolf. Road marking detection using lidar reflective intensity data and its application to vehicle localization. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 584–589. IEEE, 2014.

[32] Min Zou and Yoichi Kageyama. Road curb detection based on a deep learning framework. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0259–0262. IEEE, 2023.

[33] Lucero M Romero, Jose A Guerrero, and Gerardo Romero. Road curb detection: a historical survey. *Sensors*, 21(21):6952, 2021.

[34] 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.

[35] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.

[36] Joseph Ndong and Kavé Salamatian. Signal processing-based anomaly detection techniques: a comparative analysis. In *Proc. 2011 3rd International Conference on Evolving Internet*, pages 32–39, 2011.

[37] Yuting Chen, Jing Qian, and Venkatesh Saligrama. A new one-class svm for anomaly detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3567–3571. IEEE, 2013.

[38] Amirreza Moradi and Sajjad Asiaei Mojarad. A modified wavelet-based method for detection of outliers in time series. *Earth Observation and Geomatics Engineering*, 3(1):77–83, 2019.

[39] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[40] Dong Xu, Yanjun Wang, Yulong Meng, and Ziying Zhang. An improved data anomaly detection method based on isolation forest. In *2017 10th international symposium on computational intelligence and design (ISCID)*, volume 2, pages 287–291. IEEE, 2017.

[41] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[42] Ahmad MH Al-Khazaleh, S Al Wadi, and Faisal Ababneh. Wavelet transform asymmetric winsorized mean in detecting outlier values. *Far East Journal of Mathematical Sciences*, 96(3):339–351, 2015.

[43] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[44] Cort J Willmott. On the evaluation of model performance in physical geography. *Spatial statistics and models*, pages 443–460, 1984.

[45] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

A Calibration Table

This section represents the calibration data used for determining the pixel location of the bike in a given bike ride.

Table 2: Calibration data for each bike ride.

Bike ride Index Scale Factor	X	Y
1	577	543
2	568	553
3	602	549
4	655	514
5	523	670
6	559	571
8	557	643
11	638	496
15	573	641
17	551	649
19	599	628
20	579	658
23	561	643
24	628	636
27	603	587
28	594	668
29	603	587

B Supplementary Materials

This section presents supplementary tables referenced in Section 3, providing expanded documentation of the experimental configurations and results.

B.1 Experimental Design

Table 3 presents the experimental design used in this research. The setup follows a full factorial design with three factors: (1) the method used to acquire landmark reference segmentation masks, which determines the distance between the bike and the landmark in question; (2) the use of a rolling average—whether it is applied to smooth the final swerving signal or not; and (3) the 1D outlier detection method used to filter out potential outliers, with the options being no method, a wavelet-based method, or Isolation Forest.

Table 3: Experimental Design

Method	Rolling Average	outlier Detection
MCP2	No	Isolation Forest
MCP2	No	None
MCP2	No	Wavelet
MCP2	Yes	Isolation Forest
MCP2	Yes	None
MCP2	Yes	Wavelet
MCP3	No	Isolation Forest
MCP3	No	None
MCP3	No	Wavelet
MCP3	Yes	Isolation Forest
MCP3	Yes	None
MCP3	Yes	Wavelet
P1	No	Isolation Forest
P1	No	None
P1	No	Wavelet
P1	Yes	Isolation Forest
P1	Yes	None
P1	Yes	Wavelet

B.2 Raw Experimental Metric Overview

This section presents the experimental results obtained by applying the proposed methodology to the three datasets (A, B, and C) under the specified experimental parameters. For each dataset, performance metrics, including mean absolute error (MAE), root mean squared error (RMSE), maximum absolute error (MaxAE), and Spearman's rank correlation coefficient (Spearman's ρ), are computed and summarized in Tables 5, 6, and 7, respectively. Additionally, an aggregated evaluation across all datasets is provided in Table 4, which includes the corresponding experimental parameters alongside the performance metrics for each experiment.

Table 4: Performance metrics across method configurations across all datasets

Method	Rolling Avg.	outlier Det.	MAE ↓	RMSE ↓	Max AE ↓	Spearman's ρ ↑
MCP2	No	Isolation Forest	2.90	4.45	23.56	0.92
MCP2	No	None	4.28	9.26	78.30	0.91
MCP2	No	Wavelet	3.43	6.51	54.35	0.92
MCP2	**Yes**	**Isolation Forest**	**2.68**	**3.92**	**19.43**	**0.93**
MCP2	Yes	None	3.86	6.30	37.69	0.90
MCP2	Yes	Wavelet	3.32	5.80	35.33	0.92
MCP3	No	Isolation Forest	4.14	7.14	48.59	0.88
MCP3	No	None	5.26	10.71	83.50	0.86
MCP3	No	Wavelet	4.27	8.23	67.15	0.88
MCP3	Yes	Isolation Forest	3.72	6.22	34.72	0.87
MCP3	Yes	None	4.65	8.51	54.59	0.84
MCP3	Yes	Wavelet	4.08	7.66	48.91	0.87
P1	No	Isolation Forest	3.63	6.38	33.69	0.91
P1	No	None	6.67	16.71	133.49	0.89
P1	No	Wavelet	5.86	15.01	124.00	0.90
P1	Yes	Isolation Forest	3.34	5.42	27.95	0.92
P1	Yes	None	5.89	10.90	71.89	0.88
P1	Yes	Wavelet	5.34	10.57	71.80	0.89

Note: MAE = Mean Absolute Error (cm), RMSE = Root Mean Square Error (cm), Max AE = Maximum Absolute Error (cm), ρ = Spearman's rank correlation coefficient.

Table 5: Performance metrics across method configurations (Dataset A)

Method	Rolling Avg.	outlier Det.	MAE ↓	RMSE ↓	Max AE ↓	Spearman's ρ ↑
MCP2	No	Isolation Forest	3.60	5.27	21.53	0.91
MCP2	No	None	4.69	8.81	64.38	0.90
MCP2	No	Wavelet	4.33	7.22	51.18	0.91
MCP2	Yes	Isolation Forest	3.38	4.78	18.41	0.92
MCP2	Yes	None	4.44	6.95	37.92	0.89
MCP2	Yes	Wavelet	4.25	6.80	36.25	0.91
MCP3	No	Isolation Forest	5.52	7.72	42.26	0.84
MCP3	No	None	6.19	9.86	64.89	0.82
MCP3	No	Wavelet	5.98	9.01	55.61	0.83
MCP3	Yes	Isolation Forest	4.43	6.24	27.04	0.84
MCP3	Yes	None	4.81	7.00	34.16	0.82
MCP3	Yes	Wavelet	4.72	6.93	35.55	0.83
P1	No	Isolation Forest	3.29	4.97	26.27	0.91
P1	No	None	4.26	9.29	90.07	0.89
P1	No	Wavelet	4.18	8.87	78.62	0.89
P1	Yes	Isolation Forest	2.90	4.17	19.10	0.93
P1	Yes	None	3.65	6.63	45.18	0.89
P1	Yes	Wavelet	3.64	6.55	43.79	0.90

Note: MAE = Mean Absolute Error (cm), RMSE = Root Mean Square Error (cm), Max AE = Maximum Absolute Error (cm), ρ = Spearman's rank correlation coefficient.

Table 6: Performance metrics across method configurations (Dataset B)

Method	Rolling Avg.	outlier Det.	MAE ↓	RMSE ↓	Max AE ↓	Spearman's ρ ↑
MCP2	No	Isolation Forest	2.19	3.54	20.06	0.94
MCP2	No	None	2.89	6.97	78.22	0.93
MCP2	No	Wavelet	2.48	5.61	56.07	0.94
MCP2	Yes	Isolation Forest	2.16	3.27	16.59	0.94
MCP2	Yes	None	2.65	4.55	32.40	0.92
MCP2	Yes	Wavelet	2.41	4.19	26.05	0.93
MCP3	No	Isolation Forest	2.43	4.71	32.46	0.93
MCP3	No	None	3.35	8.18	76.76	0.91
MCP3	No	Wavelet	2.86	6.82	58.69	0.92
MCP3	Yes	Isolation Forest	2.74	5.02	26.32	0.92
MCP3	Yes	None	3.63	7.04	43.41	0.88
MCP3	Yes	Wavelet	3.23	6.43	40.10	0.90
P1	No	Isolation Forest	3.34	6.54	35.15	0.94
P1	No	None	6.07	15.95	133.72	0.94
P1	No	Wavelet	6.07	15.95	133.72	0.94
P1	Yes	Isolation Forest	3.35	6.09	30.18	0.93
P1	Yes	None	5.49	10.39	64.02	0.91
P1	Yes	Wavelet	5.49	10.39	64.02	0.91

Note: MAE = Mean Absolute Error (cm), RMSE = Root Mean Square Error (cm), Max AE = Maximum Absolute Error (cm), ρ = Spearman's rank correlation coefficient.

Table 7: Performance metrics across method configurations (Dataset C)

Method	Rolling Avg.	outlier Det.	MAE ↓	RMSE ↓	Max AE ↓	Spearman's ρ ↑
MCP2	No	Isolation Forest	2.92	4.55	29.09	0.92
MCP2	No	None	5.25	12.00	92.30	0.90
MCP2	No	Wavelet	3.49	6.72	55.80	0.92
MCP2	Yes	Isolation Forest	2.51	3.71	23.29	0.94
MCP2	Yes	None	4.50	7.39	42.74	0.88
MCP2	Yes	Wavelet	3.30	6.40	43.69	0.92
MCP3	No	Isolation Forest	4.47	8.98	71.05	0.88
MCP3	No	None	6.24	14.09	108.85	0.86
MCP3	No	Wavelet	3.97	8.87	87.16	0.90
MCP3	Yes	Isolation Forest	4.01	7.40	50.79	0.85
MCP3	Yes	None	5.50	11.49	86.19	0.81
MCP3	Yes	Wavelet	4.28	9.62	71.09	0.87
P1	No	Isolation Forest	4.25	7.65	39.65	0.88
P1	No	None	9.67	24.89	176.68	0.86
P1	No	Wavelet	7.34	20.20	159.67	0.87
P1	Yes	Isolation Forest	3.77	6.00	34.59	0.90
P1	Yes	None	8.53	15.68	106.47	0.83
P1	Yes	Wavelet	6.89	14.78	107.61	0.86

Note: MAE = Mean Absolute Error (cm), RMSE = Root Mean Square Error (cm), Max AE = Maximum Absolute Error (cm), ρ = Spearman's rank correlation coefficient.

B.3 Anova Result Tables

Given the aggregated evaluation across the entire dataset, where MAE, RMSE, Max AE, and Spearman's $\rho$are computed by comparing the measured swerving signal to the ground truth annotations for each experiment in the full factorial design, an ANOVA analysis is performed for each metric. This section presents the resulting ANOVA tables for each metric.

Table 8: ANOVA Results for MAE Across All Datasets

	Coeff.	Std.Err.	t	P> |t|	[0.025	0.975]
Intercept	4.415	0.276	15.979	1.88e-09	3.813	5.017
C(method)[T.mcp3]	0.941	0.276	3.404	0.00523	0.339	1.543
C(method)[T.p]	1.708	0.276	6.181	4.72e-05	1.106	2.310
C(rolling_avg)[T.ra]	-0.395	0.226	-1.749	0.106	-0.886	0.097
C(outlier_method)[T.isolation_forest]	-1.698	0.276	-6.143	5.00e-05	-2.300	-1.095
C(outlier_method)[T.wavelet]	-0.718	0.276	-2.598	0.0233	-1.320	-0.116
	sum_sq	**df**	**F**	**PR(>F)**		
C(method)	8.781	2	19.168	0.000184		
C(rolling_avg)	0.700	1	3.058	0.106		
C(outlier_method)	8.713	2	19.019	0.000190		
Residual	2.749	12				

Table 9: ANOVA Results for RMSE Across All Datasets

	Coeff.	Std.Err.	t	P> |t|	[0.025	0.975]
Intercept	9.182	1.015	9.046	1.05e-06	6.970	11.393
C(method)[T.mcp3]	2.038	1.015	2.008	0.068	-0.174	4.249
C(method)[T.p]	4.792	1.015	4.721	0.000496	2.581	7.004
C(rolling_avg)[T.ra]	-2.123	0.829	-2.562	0.0249	-3.929	-0.318
C(outlier_method)[T.isolation_forest]	-4.808	1.015	-4.737	0.000483	-7.019	-2.596
C(outlier_method)[T.wavelet]	-1.434	1.015	-1.412	0.183	-3.645	0.778
	sum_sq	**df**	**F**	**PR(>F)**		
C(method)	69.413	2	11.229	0.001784		
C(rolling_avg)	20.289	1	6.564	0.024907		
C(outlier_method)	73.113	2	11.827	0.001453		
Residual	37.090	12				

Table 10: ANOVA Results for Max AE Across All Datasets

	Coeff.	Std.Err.	t	P> |t|	[0.025	0.975]
Intercept	73.317	9.751	7.519	7.05e-06	52.071	94.563
C(method)[T.mcp3]	14.800	9.751	1.518	0.155	-6.446	36.046
C(method)[T.p]	35.695	9.751	3.661	0.00326	14.449	56.941
C(rolling_avg)[T.ra]	-27.147	7.962	-3.410	0.00518	-44.494	-9.800
C(outlier_method)[T.isolation_forest]	-45.252	9.751	-4.641	0.000570	-66.498	-24.006
C(outlier_method)[T.wavelet]	-9.649	9.751	-0.990	0.342	-30.895	11.596
	sum_sq	**df**	**F**	**PR(>F)**		
C(method)	3859.606	2	6.765	0.01078		
C(rolling_avg)	3316.378	1	11.626	0.00518		
C(outlier_method)	6816.707	2	11.949	0.001395		
Residual	3423.023	12				

Table 11: ANOVA Results for Spearman's ρ Across All Datasets

	Coeff.	Std.Err.	t	P> |t|	[0.025	0.975]
Intercept	0.908	0.00497	182.601	4.89e-22	0.897	0.919
C(method)[T.mcp3]	-0.0514	0.00497	-10.341	2.49e-07	-0.0623	-0.0406
C(method)[T.p]	-0.0199	0.00497	-3.993	0.00179	-0.0307	-0.00902
C(rolling_avg)[T.ra]	-0.00860	0.00406	-2.119	0.0556	-0.0175	0.000244
C(outlier_method)[T.isolation_forest]	0.0271	0.00497	5.453	0.000147	0.0163	0.0380
C(outlier_method)[T.wavelet]	0.0169	0.00497	3.398	0.00529	0.00607	0.0277
	sum_sq	**df**	**F**	**PR(>F)**		
C(method)	0.00807	2	54.390	9.62e-07		
C(rolling_avg)	0.000333	1	4.490	0.0556		
C(outlier_method)	0.00225	2	15.167	0.000519		
Residual	0.000891	12				