Near-incident detection in railroad environments: lateral distance estimation from train-mounted monocular camera

Yilei Wang, Giacomo D'Amicantonio, Egor Bondarev
Eindhoven University of Technology

Code: https://github.com/wwwyilei/RailSafety

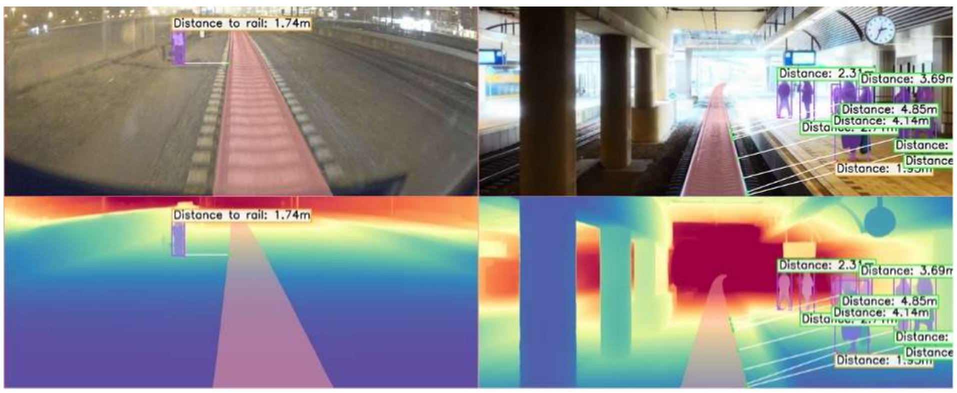

Figure 1. NMD performance across diverse scenarios. Each pair shows the original image with distance measurements (top) and corresponding depth visualization (bottom). Left: pedestrian detection and distance measurement in low-light conditions. Right: Multi-person detection and distance estimation in crowded, poorly-lit platform environments.

Abstract

With the increasing number of incidents in complex railroad environments, there is an urgent need for automated systems that can predict dangerous events through accurate detection and distance measurement between trains and various hazards. To address these challenges, we introduce Near-Miss Detector (NMD), an integrated framework that leverages specialized models to accurately detect possible collisions between trains and people or objects in the railroad environment via monocular cameras installed in front of a train. NMD constructs a comprehensive, 3-dimensional view of a given scene via object detection, instance segmentation and depth estimation. In this view, the risk of an accident is measured through the distance between the moving train and detected objects in the scene. In order to apply NMD in real-world scenarios, we present a novel depth-calibration mechanism based on constant geometrical properties of a railroad environment, such as the gauge of the rail track. To validate our work, we collected a dataset of measurements from multiple train stations in order to accurately represent the diversity and challenges of a complex railroad environment. NMD demonstrates robust performance in object detection, track segmentation, and distance measurement while maintaining suitable processing latency. This work contributes to the field of automated railway safety monitoring by showing the feasibility of monocular vision-based distance measurement in complex railway environments, offering a cost-effective solution for improving railway safety systems.

1. Introduction

In modern transportation systems, railways serve as a critical infrastructure component, enabling efficient movement of people and goods while supporting economic and social activities. However, the dense use of railway networks presents significant safety challenges, particularly evident in the Netherlands [25]. Recent statistics from 2023 indicate the severity of these challenges, as the Dutch railway system recorded 200 serious incidents: 84 at level crossings, 31 at train stations, and 85 on open tracks [2, 26].

Consequently, railway operators have begun to gradually deploy video surveillance systems to enhance monitoring of potential threats. However, current implementations predominantly rely on fixed camera installations at specific locations, such as level crossings, focusing on predictable scenarios and regular human activity [1, 4, 12]. Many crucial areas remain unmonitored, such as station platforms, open tracks between stations, and railway yards, severely restricting the detection and response capabilities for dynamic threats and unexpected events. To address these shortcomings, recent research has explored various monitoring approaches. Multi-sensor systems incorporating LiDAR, radar, or stereo cameras provide accurate distance measurements, but at high costs and complex integration requirements [3, 34]. Monocular camera-based systems, particularly forward-facing cameras mounted on trains, offer a more practical and cost-effective alternative for continuous surveillance during operation [10].

Currently, train operators are using monocular camera systems to monitor, analyze, and quantify near-miss events. A near-miss is typically defined as a situation where the driver of a moving train takes emergency actions such as audible warning and/or brake application [24] to avoid a collision. The decision of what constitutes a near-miss situation is typically left to the driver. These events are more frequent than actual collisions and provide valuable insight into safety performance and risk analysis. NMD provides a framework to automatically detect when an object is within dangerous range from the moving train by estimating the distance between the object and the train, called "lateral distance". Furthermore, NMD allows for train operator to unify the situations in which a near-miss is detected, reducing the impact of the driver's decision-making.

NMD is an integrated method that leverages of recent advances in deep learning and computer vision. The method combines four state-of-the-art models in a parallel processing architecture. YOLO-World [6] for robust hazard identification, Edge-SAM [40] for precise instance segmentation, TEP-Net [20] for accurate rail boundary prediction and Depth-Anything V2 [37] for reliable distance estimation. To address the fundamental challenge of distance measurement, NMD combines depth-guided point matching with track gauge scale calibration, enabling accurate distance estimation without additional sensors. To validate our approach and address the lack of datasets with lateral distance measurements, we construct a specialized dataset with calibrated distance annotations in complex train station environments. NMD is implemented through asynchronous execution streams, enabling efficient processing while maintaining high accuracy.

The main contributions of this work are the following:

(a) A measurement method for lateral distance between rail-track and scene objects that enables reliable monocular distance estimation by combining depth-guided point matching with track width scale calibration.
(b) A novel two-stage calibration method that integrates near-field sleeper-based calibration with far-field geometric constraints from rail tracks to convert relative depth estimates into real-world depth measurement.
(c) A railway station dataset with accurate lateral distance annotations, including annotations for tracks, persons, and vehicles.

2. Related Work

This section reviews related work in forward camera-based railway safety monitoring, focusing on five key technical aspects: object detection, instance segmentation, track detection, monocular depth estimation and lateral distance measurement methods. Recent studies have explored near-miss detection in railway environments with direct approaches. For example, [1] focused only on vehicle detection within predefined regions and utilized train operation data, such as braking and horn signals to identify near-miss events. On the other hand, the authors of [8] utilized track segmentation and obstacle detection to analyze the pixel relationship between their edges, establishing an assessment method based on this analysis. However, both methods failed to achieve accurate lateral distance computations.

2.1. Object Detection in Railway Environments

Early methods developed for object detection in railway environments relied on appearance-based techniques, such as template matching [11] and motion-based detection through frame differencing [18] or optical flow analysis [17]. Although these approaches provided fundamental solutions, they struggled with environmental variations and complex scenes typical for railway operations. More recent, deep-learning based approaches such as two-stage detectors [13, 16], achieve high detection precision through sequential region proposal and classification steps. In railway applications, He et al. [16] demonstrated the effectiveness of this approach by adapting R-CNN variants for railway component detection, achieving detection accuracies exceeding 90%. However, the computational overhead of these two-stage architectures presents significant challenges for real-time deployment in safety-critical railway systems. Single-stage detectors like YOLO [27] addressed these limitations by performing localization and classification simultaneously in a single network pass. This architectural design substantially reduces computational complexity while maintaining competitive accuracy. YOLO-World [6] further advanced this approach by incorporating zero-shot learning capabilities, enabling the detection of previously unseen object categories without additional training. This is particularly valuable in railway safety applications, where the system should rapidly identify and respond to diverse and potentially unexpected hazards in varying operational conditions.

2.2. Instance Segmentation for Precise Object Delineation

In railway safety applications, accurate analysis of spatial relationships between objects and tracks demands precise object boundary information, which cannot be adequately provided by the boundary box detection alone. Instance segmentation addresses this limitation by enabling exact boundary detection [23]. Early approaches, such as [5, 22, 28], introduced the concept of end-to-end semantic segmentation but struggled with precise boundary detection, requiring enhanced feature extraction and dedicated boundary refinement mechanisms. However, these methods often required extensive task-specific training and struggled with real-time performance requirements. Recent developments in foundation models have created new opportunities for railway safety applications. The Segment Anything Model (SAM) [19] introduced zero-shot generalization capabilities, enabling robust segmentation of diverse objects without domain-specific training. EdgeSAM [40] further optimized this approach for real-time applications through efficient edge detection and streamlined architecture, making it particularly relevant for dynamic railway safety scenarios where rapid processing is essential.

2.3. Rail Track Detection and Segmentation

Early research on rail track detection and segmentation relied on traditional computer vision techniques, such as template matching [11] and gradient-based edge detection [31]. However, these methods struggled with variations in track shapes and environmental conditions. Deep learning approaches focusing on semantic segmentation, such as RailNet [35], introduced multi-scale feature aggregation to manage diverse track appearances and environmental variations, showing improved robustness over traditional methods. However, these segmentation-based approaches faced a limitation in reliably distinguishing the ego-track from adjacent tracks, especially at complex junctions and intersections where multiple tracks converge. Recent research has shifted towards regression-based approaches for more precise track localization. TEP-Net [20] formulated the track detection task as a coordinate regression problem rather than pixel-wise segmentation. By directly predicting track boundary coordinates, this approach achieved superior performance in ego-track identification. This approach was particularly effective in complex scenarios with multiple parallel tracks and varying lighting condition.

2.4. Monocular Depth Estimation in Railway Environments

Depth estimation in railway safety systems has evolved from active sensor-based approaches to passive vision-based solutions. While active sensors like LiDAR provide precise measurements, their high costs limit practical deployment [9, 39]. This has driven increasing research interest towards monocular vision-based solutions. Early approaches relied on geometric methods that leverage scene structure and camera parameters. These methods were adapted from vehicle-focused techniques, such as those proposed in [33], which estimate distances using known object dimensions and camera parameters. The authors of [21] extended these approaches by developing inverse perspective mapping techniques for railway scenes. However, these geometric methods struggled with complex platform geometries and varying lighting conditions typical in railway environments. Deep learning approaches to depth estimation in railway environments face challenges due to the lack of specific railway depth datasets, despite considerable progress in methods developed for city scenes. However, several monocular depth estimation techniques have a potential for railway applications. Self-supervised methods [14, 41] have proposed depth estimation from unlabeled video sequences using photometric reconstruction. Similarly, advances in temporal consistency. [15, 36] provided insight for maintaining stable depth estimates in railway safety monitoring. The emergence of foundational models such as Depth Anything [37] lead to significant advancements due to their strong generalization capability, making it suitable for railway environments.

2.5. Lateral Distance Measurement

Accurate measurement of lateral distances between railway tracks and potential obstacles is crucial for near-miss detection and risk assessment. Traditional approaches have primarily relied on stereo vision systems [29] or perspective transformation techniques [1]. While stereo systems can provide accurate measurements, they require precise camera calibration and face reliability issues in varying environmental conditions. Although perspective transformation methods are simpler to implement, they often struggle due to the complexities of railway environments such as differences in platform height.

3. Method

The proposed NMD framework, illustrated in Fig. 2, integrates four state-of-the-art deep learning models into a parallel processing pipeline to provide essential scene understanding and depth information. NMD consists of three parallel processing streams that feed into an integrated lateral distance measurement module. The object analysis stream combines YOLO-World for object detection and EdgeSAM for instance segmentation. The track analysis stream employs TEP-Net for track boundary detection and mask generation. The depth estimation stream utilizes Depth Anything V2 with a custom calibration process for real-world depth prediction. The lateral distance measurement module integrates these outputs by first identifying key points

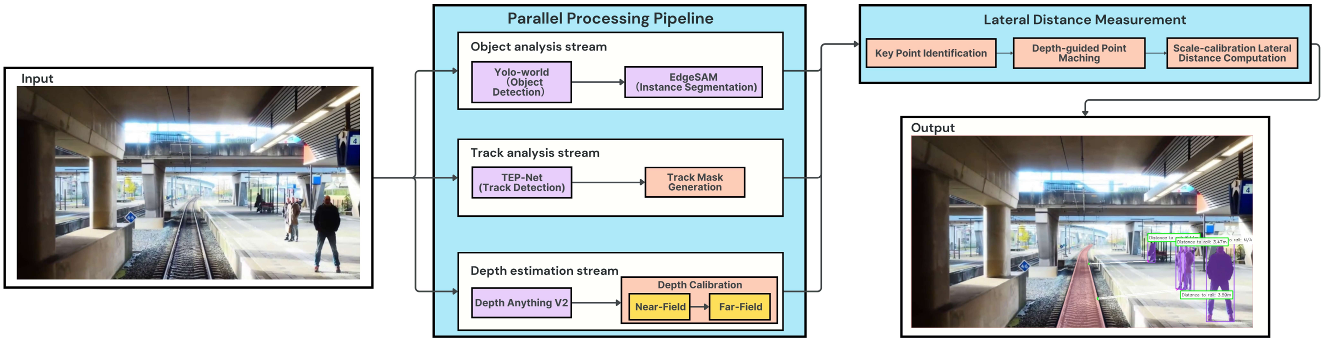

Figure 2. Overview of the Near-Miss Detection framework. NMD consists of three parallel processing streams integrated with a lateral distance measurement module. Each stream feed into the lateral distance measurement module that performs key point identification, depth-guided point matching, and scale-calibrated distance computation for safety assessment. In orange the contributions of this work.

in the objects, then performing depth-guided point matching between the objects and tracks, and finally computing the lateral distance using the scale calibration based on the standard track gauge (1.435m).

Object Analysis Stream. Railway safety monitoring requires accurate object detection in the scene. In the object detection and instance segmentation stream, YOLO-World [6] and EdgeSAM [40] are combined in a two-stage pipeline that identifies possible factors of risk in the scene, such as people or vehicles. The bounding boxes produced by YOLO-World are refined by EdgeSAM into accurate segmentation masks, providing NMD the capability of correctly identify pixels belonging to objects that represent possible risks. Both models, being trained on large-scale datasets, are robust to different scenes, light condition and other environmental factors.

Track Analysis Stream. Accurate rail track detection presents challenges in railway safety systems, particularly in distinguishing the track currently being used by the train, called "ego-track", from adjacent tracks while maintaining robustness to environmental variations. To address this issue, TEP-Net [20] allows NMD to detect the boundaries of a railway track instead of relying on less accurate semantic segmentation approaches. TEP-Net dynamically adjusts the detection region through time, effectively distinguishing the ego-track from adjacent and overlapping tracks in different environmental conditions.

Depth Estimation Stream. In order to build an accurate, tri-dimensional representation of the scene, NMD employs Depth Anything V2 (DAV2) [37] to obtain the relative depth of each pixel in the scene via zero-shot depth estimation. Trained on large-scale datasets, DAV2 is a robust choice for this stream, estimating accurate depth values in topologically complex and crowded scenes. However, these estimates reflect only relative depth relationships without absolute metric scale, necessitating an effective calibration approach. To this end, we propose a two-stage calibration method that combines near-field and far-field depth refinement, utilizing the standard sleeper spacing for initial depth calibration and then the rail track geometry as a robust reference.

Near-field depth calibration. The primary calibration stage uses the standard sleeper spacing as a reference metric, where manually annotated sleeper positions establish the relationship between relative depths and real-world depth measurements. We then fit an exponential function to the obtained measurements, establishing a non-linear relationship between distances in the pixel space and real-world coordinates. The calibration function obtained is shown in Equation 1:

$$D_{calibrated} = a \cdot e^{bD_{raw}} + c, \tag{1}$$

where $D_{calibrated}$ is the calibrated depth value, D_{raw} is the raw depth estimate from DAV2, and parameters a, b, and c are optimized through curve fitting using known sleeper spacing measurements.

Far-field depth calibration. A second calibration stage employs the rail tracks as geometric references. Rail tracks maintain distinct boundaries and a consistent width (1.435m) in the real world, which are both leveraged as geometric constraints. **(I)** In railway scenes, parallel rail tracks appear to converge towards a vanishing point due to perspective projection principles. By utilizing the vanishing point (vp_y), as illustrated in Fig. 3, a depth scaling relationship can be established, shown in Equation 2:

$$\frac{D(y)}{D_{bottom}} = \frac{vp_y - y}{vp_y - bottom_y}, \tag{2}$$

where $D(y)$ is the depth to be estimated at position y, D_{bottom} is the known depth at the reference position $bottom_y$, vp_y is the vertical coordinate of the vanishing point, y is the vertical coordinate of the measured point, and $bottom_y$ is the vertical coordinate of the reference point close to the camera.

(II) The second geometric constraint utilizes the principle of similar triangles formed by the optical center of the

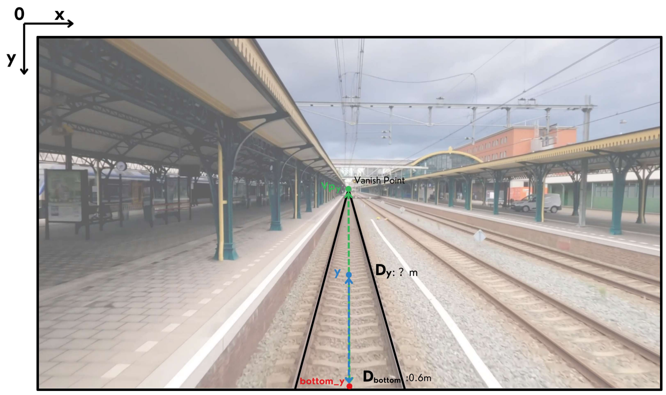

Figure 3. Perspective-based depth scaling. The D_{bottom} distance is set based on the sleepers spacing, while the vanishing point can be detected.

camera and the rail track at different depths. Since the actual rail gauge width remains constant (1.435m) in the real-world, as illustrated in Fig. 4, any apparent width changes in the image plane are solely due to the viewing distance. This relationship can be expressed as shown in Equation 3:

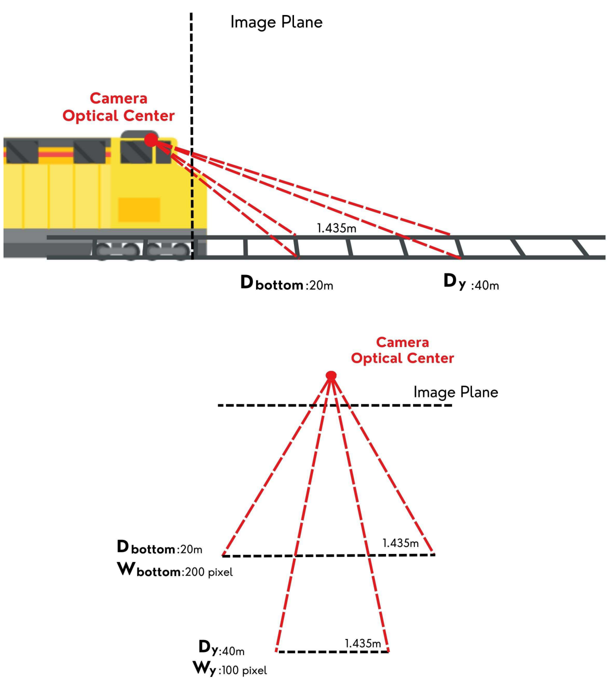

Figure 4. Rail width consistency between different projective points.

$$\frac{D(y)}{D_{bottom}} = \frac{W_{bottom}}{W(y)}, \tag{3}$$

where $D(y)$ is the depth to be estimated at position y, D_{bottom} is the known depth at the reference point, W_{bottom} is the rail width (in pixels) at a reference point close to the

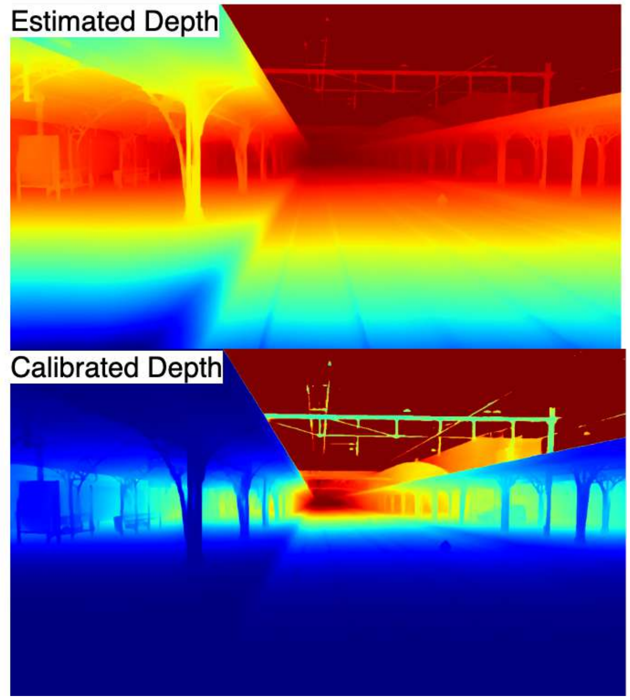

Figure 5. Comparison between depth values estimated by DAV2 before and after the proposed calibration process. Top: Estimated depth map. Bottom: Calibrated depth map.

camera, $W(y)$ is the measured rail width at vertical position y. By measuring the rail width in pixels at different positions and using the sleeper-calibrated reference point, the relative depths can be estimated through this inverse proportional relationship. The geometric-based depth $D(y)$ combines these two constraints in Equation 4:

$$D(y) = D_{bottom} \cdot \left(\left(\frac{vp_y - y}{vp_y - bottom_y} \right) + \left(\frac{W_{bottom}}{W(y)} \right) \right) \tag{4}$$

To smoothly integrate both calibration methods, a distance-adaptive weighting mechanism was implemented as shown in Equation 5:

$$D_{optimized} = D_{calibrated} + w(y) \cdot \delta(y) \tag{5}$$

where $D_{optimized}$ is the final optimized depth value, $D_{calibrated}$ is the depth from the primary calibration stage, $\delta(y)$ is the difference between geometric-based and sleeper-based estimates, and $w(y)$ is the weight factor that dynamically adjusts based on distance. The weight $w(y)$ follows a sigmoid function to ensure a smooth transition between the two calibration methods, as shown in Equation 6:

$$w(y) = \frac{1}{1 + e^{-s(d_{norm} - 0.3)}} \tag{6}$$

where s is the steepness parameter controlling the transition rate, d_{norm} is the normalized distance from the camera

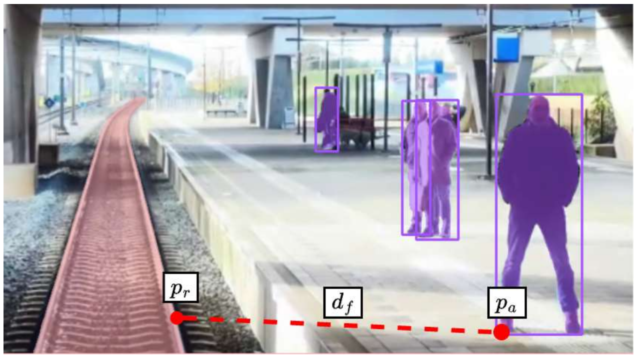

Figure 6. Visualization of the pixels selected by NMD and the distance between them.

(ranging from 0 to 1), and 0.3 denotes the transition starting point, set at 30% of the track length from the camera.

This sigmoid-based weighting prioritizes sleeper-based calibration near the camera and gradually increases the influence of geometric constraints as the distance increases. As illustrated in Fig. 5, the calibration method makes depth relationships more consistent with real-world appearances in the image.

Lateral Distance Measurement. At this stage, the results of the three streams are combined to measure the lateral distance between an object and the moving train. The object analysis stream allows NMD to identify a set of pixels P_a that constitute the semantic mask of an object, while the track analysis stream produces the set of pixels P_r of the ego-track. NMD then selects a set of candidate pixels $P_{ac} \in P_a$ belonging to an object and a set of pixels $P_{rc} \in P_r$ belonging to the ego-track by comparing their depth values $D(p_a)$ and $D(p_r)$ obtained by the depth estimation stream:

$$(P_{ac}, P_{rc}) = |D(p_i) - D(p_j)| < \epsilon, \forall p_i \in P_a, p_j \in P_r, \tag{7}$$

where $\epsilon = 0.2$ is a threshold that accounts for small inaccuracies in the estimated depth. From P_{ac} and P_{rc}, NMD selects the pair of points p_a, p_r, as shown in Fig. 6, that are closest to each other via:

$$d = min_{p_a \in P_{ac}, p_r \in P_{rc}} \sqrt{(x_{p_a} - x_{p_r})^2 + (y_{p_a} - y_{p_r})^2}. \tag{8}$$

To account for perspective distortion, we introduce a scaling factor s based on the standard railway track gauge:

$$s = \frac{1.435}{max(p_{i_x}) - min(p_{i_x})}, \forall p_i \in P_{rc} | D(p_i) = D(p_r), \tag{9}$$

where the denominator is the distance between the boundaries of the track at the depth of pixel p_r along the image's x-axis, while the numerator is the standard gauge of railway tracks. The final lateral distance measurement d_f is then calculated as:

$$d_f = d * s, \tag{10}$$

4. Experimental Results

Experimental Setup Existing railway datasets, such as RailSem19 [38], RAWPED [32], RailSet [42], and OSDaR23 [30], primarily focus on object detection and segmentation tasks. However, they lack precise ground-truth lateral distance annotations, making them unsuitable for validation of monocular lateral distance measurement method, as summarized in Table 1. We constructed a specialized validation dataset consisting of 680 high-resolution images (1920×1080 pixels) extracted from publicly available train journey videos. The dataset includes 2058 pedestrian detection annotations, 20 vehicle detection annotations, and 680 rail masks. To ensure accurate lateral distance measurements, field measurements were conducted at multiple railway stations. These measurements revealed consistent infrastructure dimensions that serve as reliable physical references, as shown in Fig. 7. The key measured distances include:

- Track edge to platform edge: 0.8m
- Platform edge to brick patterns: 0.9m
- Pole to brick spacing: 0.4m
- Pole to tactile paving: 2.0m

Dataset	Year	Seg.	Det.	Depth	Lateral.	Classes
RailSem19 [38]	2019	✓	✓	55	55	19
RAWPED [32]	2023	55	✓	55	55	1+
RailSet [42]	2023	✓	✓	55	55	5+
OSDaR23 [30]	2023	✓	✓	✓	55	20+
Ours(Validation)	2024	✓	✓	55	✓	2

Table 1. Comparison of railway safety datasets.

To ensure comprehensive dataset annotations for accurate validation, a structured annotation process was implemented, including object detection, track segmentation, and lateral distance calibration.

Object Detection and Track Segmentation Annotation. Object and track annotations were performed using a semi-automatic process. Initially, object detection bounding boxes were generated using the pre-trained YOLO-World [6] model, and rail track segmentation masks were obtained via TEP-Net [20]. Both results were manually refined using CVAT [7] to ensure accuracy, particularly for objects near the track and track boundaries at platform intersections and curved sections.

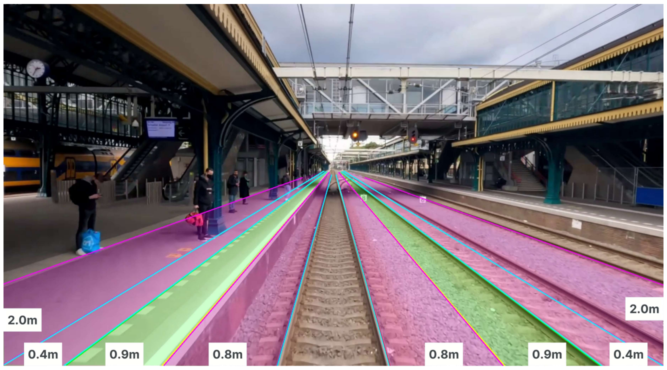

Figure 7. Lateral distance annotations of key regions in a train stations. The measurements of these regions are fixed and therefore are considered them as ground truths to evaluate NMD.

Lateral Distance Annotation. To achieve precise lateral distance annotation, a region-based calibration method was developed using the reference measurements mentioned above, as shown in Fig. 7. Each image was divided into non-overlapping zones, with homography projection matrices computed from reference features to ensure accurate coordinate mapping. The calibration process utilized the known reference distances to correct for perspective distortion and convert image coordinates to real-world values. The 3×3 transformation matrix is shown in Eq. 11:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H_r \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{11}$$

where (x', y') represents the pixel coordinates in the image, and (x, y) represents the corresponding real-world coordinates in meters. This transformation allows us to convert any point in the image to its real-world position using the calibrated homography matrix H_r. To enhance measurement consistency, the track centerline was used as a reference, and distances to both track edges were calculated symmetrically.

To validate our approach, an interactive visualization tool was developed that allows distances to be measured by clicking on points in the image. The tool displays the reference regions and real-time distance calculations, helping to verify the accuracy of the measurements.

4.1. System Performance Analysis

Object Detection Evaluation .

The evaluation focuses on the performance of person and vehicle detection in railway station environments. Two state-of-the-art zero-shot object detection models, YOLO-World and Grounding DINO, are compared. Both models can detect novel object categories without domain-specific training. Based on preliminary experiments, the models were configured with class-specific confidence thresholds: 0.19 for person detection to maximize recall of safety-critical objects, 0.4 for vehicle detection to mitigate the impact of uncertainty and reduce potential false positives. The results of the comparison are presented in Table 2.

Method	Person		Vehicle		FPS
	Precision	Recall	Precision	Recall	
YOLO-World	0.897	**0.6186**	**0.02**	**0.700**	**17.3**
Grounding DINO	**0.981**	0.4762	0.014	0.500	1.5

Table 2. Comparison of the performance of YOLO-World and Grounding DINO for person and vehicle detection.

For person detection, while Grounding DINO achieves higher precision, YOLO-World demonstrates better overall performance with higher recall and significantly faster throughput. For vehicle detection, both models show lower precision, with YOLO-World achieving a better recall rate. Given its superior recall and processing efficiency, YOLO-World was selected as our primary detection model.

Track Segmentation Evaluation Track detection performance was evaluated using the mIoU and inference throughput metrics. Table 3 compares the regression-based TEP-Net approach with a semantic segmentation baseline that uses EdgeSAM with manually points inputs as prompt.

Method	IoU	FPS
TEP-Net	**0.974**	16.5
Semantic Segmentation	0.839	**17**

Table 3. Comparison of track segmentation methods. TEP-Net achieves a significantly higher IoU while processing slightly less frames per second.

TEP-Net demonstrates superior segmentation accuracy, achieving a 97.4% IoU compared to 83.9% for the semantic segmentation baseline. While the inference times are comparable, TEP-Net's higher accuracy in boundary delineation is crucial for precise lateral distance measurements.

Lateral Distance Measurement Evaluation. The analysis of 1,156 valid measurements reveals our system's lateral distance measurement performance. As shown in Table 4, compared to ground truth, the majority of estimated measurements are below the error threshold of 0.3 meters, with 8.2% being closer than 10cm to the ground-truth.

Additionally, as shown in Table 5, 67% of the measurements fall within the critical safety range of 2-3 meters from the track.

Larger errors (>0.5 in 9.2% of cases) are primarily observed at the boundaries of the validated measurement range or in regions with significant perspective distortion. These scenarios often pose challenges for depth estimation and homography-based distance calculations, suggest-

Error Range	Number of Cases	Percentage (%)
≤ 0.1m	95	8.2%
0.1m - 0.3m	592	**51.2**%
0.3m - 0.5m	363	31.4%
>0.5m	106	9.2%

Table 4. Distribution of the lateral distance measurement errors. The majority of errors are below the 30cm threshold, and more than 90% of errors are lower than 50cm.

Distance Range	Number of Cases	Percentage (%)
0m - 1m	23	2.0%
1m - 2m	227	19.6%
2m - 3m	774	**67.0**%
3m - 4m	132	11.4%

Table 5. Distribution of lateral distance ground-truths measured in different train stations.

ing the need for enhanced calibration techniques in future work.

Depth Estimation and Calibration Analysis Comprehensive ablation studies evaluate the effectiveness of the depth-guided approach and its calibration component, with results presented in Table 6.

Method	Error Metrics (m)			Error Dist. (%)		
	Mean	Std	Max	≤ 0.3	≤ 0.5	>0.5
NMD w/o depth	1.262	3.808	74.650	30.2	15.6	54.2
NMD w/o calibration	0.375	2.154	51.919	**59.5**	**31.5**	**9.0**
NMD	**0.290**	**0.195**	**2.581**	59.4	31.4	9.2

Table 6. Impact of depth prediction and calibration on absolute errors and distribution of errors across space scales.

The ablation results demonstrate the effectiveness of the depth-guided method, which achieves a mean error of 0.290m and maintains notably low variance with a standard deviation of 0.195m and maximum error of 2.581m. Notably, 90.6% of measurements have errors below 0.5m, with 59.4% within 0.3m. Calibration significantly reduces variance and maximum error compared to the uncalibrated system. While the uncalibrated method shows a similar error distribution (59.5% $\leq$ 0.3m, 31.5% between 0.3m and 0.5m), its standard deviation increases to 2.154m, with a maximum error of 51.919m.

The importance of depth guidance is demonstrated by the sharp performance drop when depth estimation is removed. Without depth estimation, the mean error rises to 1.262m, with a standard deviation of 3.808m and a maximum error of 74.650m. Additionally, the proportion of measurements within 0.3m drops to 30.2%, while 54.2% exceed 0.5m.

System Efficiency. To evaluate the processing performance, both parallel and sequential architectures were tested on our validation dataset comprising 680 images. The parallel processing implementation achieves 2.51 FPS compared to 2.23 FPS for sequential processing. As shown in Table 7, this represents an 11.2% improvement in processing efficiency, reducing the per frame processing time from 0.449 seconds to 0.399 seconds.

Strategy	FPS	Depth estimation (s)
Parallel	**0.399**	0.304
Sequential	0.449	**0.266**

Table 7. Performance comparison of different processing strategies.

The efficiency gain from parallel processing is relatively small, leading to further investigation of individual components. Subsequent analysis revealed that the depth estimation component performs worse in the parallel setup (0.304s) compared to sequential processing (0.266s). This unexpected result is likely due to GPU-related constraints: the GPU (NVIDIA RTX 4060-8GB) operates at full memory capacity in both parallel and sequential setups, which means that the DAv2 computational structure may not fully benefit from parallelism.

5. Conclusion

This paper introduces a novel railway safety monitoring method, named NMD, that leverages deep learning and monocular vision techinques for precise lateral distance measurement. NMD allows to detect near-incident situations without requiring manual intervetions, leading to a unified definition of what constitutes a near-incident situation and allowing for further optimizations from railway operators. The key innovation is a depth-guided method that leverages track gauge (1.435m) as a dynamic scale reference, eliminating the need for precise camera calibration or extra sensors. Ablation studies show a mean error reduction from 1.262m to 0.290m, with 59.4% of measurements within 0.3m. The system's accuracy across varied conditions ensures reliability in real-world railway environments, including crowded platforms and occlusions. This practical, monocular vision-based solution enhances railway safety without significant infrastructure modifications.

References

[1] Sina Aminmansour, Frederic Maire, Gregoire S Larue, and Christian Wullems. Improving near-miss event detection rate at railway level crossings. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2015. 1, 2, 3

[2] Arriva. Met de trein. 1

[3] Amanda Berg, Kristoffer Öfjäll, Jörgen Ahlberg, and Michael Felsberg. Detecting rails and obstacles using a train-

mounted thermal camera. In *Image Analysis: 19th Scandinavian Conference, SCIA 2015, Copenhagen, Denmark, June 15-17, 2015. Proceedings 19*, pages 492–503. Springer, 2015. 2

[4] Zhiwei Cao, Yong Qin, Limin Jia, Zhengyu Xie, Yang Gao, Yaguan Wang, Ping Li, and Zujun Yu. Railway intrusion detection based on machine vision: A survey, challenges, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 1

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2, 4, 6

[7] CVAT.ai. Cvat - computer vision annotation tool. https://github.com/cvat-ai/cvat, 2024. Accessed: 2024-01-13. 6

[8] Amartuvshin Dagvasumberel, Bilguunmaa Myagmardulam, Byambasuren Myagmar, Byambasuren Luvsankhuu, and Tadachika Nakayama. Railroad near-miss occurrence detection and risk estimation system with data from camera using deep learning. In *2021 5th International Conference on Imaging, Signal Processing and Communications (ICISPC)*, pages 83–87. IEEE, 2021. 2

[9] Meng Ding, Zhenzhen Zhang, Xinyan Jiang, and Yunfeng Cao. Vision-based distance measurement in advanced driving assistance systems. *Applied Sciences*, 10(20):7276, 2020. 3

[10] Xuewen Ding, Xinnan Cai, Ziyi Zhang, Wenyan Liu, and Wenwen Song. Railway foreign object intrusion detection based on deep learning. In *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pages 735–739. IEEE, 2022. 2

[11] Jorge Corsino Espino, Bogdan Stanciulescu, and Philippe Forin. Rail and turnout detection using gradient information and template matching. In *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings*, pages 233–238. IEEE, 2013. 2, 3

[12] Muhammad Asad Bilal Fayyaz and Christopher Johnson. Object detection at level crossing using deep learning. *Micromachines*, 11(12):1055, 2020. 1

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 3

[15] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 160–170, 2022. 3

[16] Deqiang He, Ruochen Ren, Kai Li, Zhiheng Zou, Rui Ma, Yuliang Qin, and Weifeng Yang. Urban rail transit obstacle detection based on improved r-cnn. *Measurement*, 196: 111277, 2022. 2

[17] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[18] Ramesh Jain and H-H Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE transactions on pattern analysis and machine intelligence*, (2):206–214, 1979. 2

[19] 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. 3

[20] Thomas Laurent. Train ego-path detection on railway tracks using end-to-end deep learning. *arXiv preprint arXiv:2403.13094*, 2024. 2, 3, 4, 6

[21] Liang-Chien Liu, Chiung-Yao Fang, and Sei-Wang Chen. A novel distance estimation method leading a forward collision avoidance assist system for vehicles on highways. *IEEE Transactions on Intelligent Transportation Systems*, 18(4): 937–949, 2016. 3

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[23] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 3

[24] Rail Safety Regulator's Panel. Guideline for the top event classification of notifiable occurrences: Occurrence classification-guideline one (ocg1). *Rail Safety Regulator's Panel, Tech. Rep*, 2008. 2

[25] ProRail. Jaarverslag 2023, 2023. Accessed: 2023-05-23. 1

[26] ProRail. Veiligheid, 2023. Accessed: 2023-05-23. 1

[27] 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

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[29] J Song and J Yang. 1.1 distance measurement for intrusive objects on railway tracks. *Proceedings-ettc2020*, pages 8–30, 2020. 3

[30] Rustam Tagiew, Pavel Klasek, Roman Tilly, Martin Köppel, Patrick Denzler, Philipp Neumaier, Tobias Klockau, Martin Boekhoff, and Karsten Schwalbe. Osdar23: Open sensor

data for rail 2023. In *2023 8th International Conference on Robotics and Automation Engineering (ICRAE)*, pages 270–276. IEEE, 2023. 6

[31] Canan Taştimur, Mehmet Karaköse, and Erhan Akın. A vision based condition monitoring approach for rail switch and level crossing using hierarchical svm in railways. *International Journal of Applied Mathematics Electronics and Computers*, (Special Issue-1):319–325, 2016. 3

[32] Tugce Toprak, Burak Belenlioglu, Burak Aydın, Cuneyt Guzelis, and M Alper Selver. Conditional weighted ensemble of transferred models for camera based onboard pedestrian detection in railway driver support systems. *IEEE Transactions on Vehicular Technology*, 69(5):5041–5054, 2020. 6

[33] Vo Thi Bich Tram and Myungsik Yoo. Vehicle-to-vehicle distance estimation using a low-resolution camera based on visible light communications. *IEEE Access*, 6:4521–4527, 2018. 3

[34] Masato Ukai, Bogdan Tomoyuki Nassu, Nozomi Nagamine, Masato Watanabe, and Takayuki Inaba. Obstacle detection on railway track by fusing radar and image sensor. In *Proceedings of the 9th World Congress on Railway Research (WCRR), Lille, France*, pages 22–26, 2011. 2

[35] Yin Wang, Lide Wang, Yu Hen Hu, and Ji Qiu. Railnet: A segmentation network for railroad detection. *IEEE Access*, 7:143772–143779, 2019. 3

[36] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 3

[37] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2, 3, 4

[38] Oliver Zendel, Markus Murschitz, Marcel Zeilinger, Daniel Steininger, Sara Abbasi, and Csaba Beleznai. Railsem19: A dataset for semantic rail scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 6

[39] Junning Zhang, Qunxing Su, Cheng Wang, and Hongqiang Gu. Monocular 3d vehicle detection with multi-instance depth and geometry reasoning for autonomous driving. *Neurocomputing*, 403:182–192, 2020. 3

[40] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*, 2023. 2, 3, 4

[41] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3

[42] Arij Zouaoui, Ankur Mahtani, Mohamed Amine Hadded, Sébastien Ambellouis, Jacques Boonaert, and Hazem Wannous. Railset: A unique dataset for railway anomaly detection. In *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*, pages 1–6. IEEE, 2022. 6