vehicles

MDPI

Article

Real-Time 3D Scene Understanding for Road Safety: Depth Estimation and Object Detection for Autonomous Vehicle Awareness

Marcel Simeonov [1], Andrei Kurdiumov [2] and Milan Dado [1,*]

[1] Department of Multimedia and Information-Communication Technology, University of Žilina, 010 26 Žilina, Slovakia; simeonov1@stud.uniza.sk
[2] Department of Video & Image Analytics, Kapsch TrafficCom AG, 1120 Vienna, Austria; andrei.kurdiumov@kapsch.net
* Correspondence: milan.dado@uniza.sk

Abstract

Accurate depth perception is vital for autonomous driving and roadside monitoring. Traditional stereo vision methods are cost-effective but often fail under challenging conditions such as low texture, reflections, or complex lighting. This work presents a perception pipeline built around FoundationStereo, a Transformer-based stereo depth estimation model. At low resolutions, FoundationStereo achieves real-time performance (up to 26 FPS) on embedded platforms like NVIDIA Jetson AGX Orin with TensorRT acceleration and power-of-two input sizes, enabling deployment in roadside cameras and in-vehicle systems. For Full HD stereo pairs, the same model delivers dense and precise environmental scans, complementing LiDAR while maintaining a high level of accuracy. YOLO11 object detection and segmentation is deployed in parallel for object extraction. Detected objects are removed from depth maps generated by FoundationStereo prior to point cloud generation, producing cleaner 3D reconstructions of the environment. This approach demonstrates that advanced stereo networks can operate efficiently on embedded hardware. Rather than replacing LiDAR or radar, it complements existing sensors by providing dense depth maps in situations where other sensors may be limited. By improving depth completeness, robustness, and enabling filtered point clouds, the proposed system supports safer navigation, collision avoidance, and scalable roadside infrastructure scanning for autonomous mobility.

Keywords: stereo depth estimation; autonomous vehicles; real-time perception; object detection; point cloud generation; embedded systems; depth-aware navigation; roadside monitoring

Check for updates

Academic Editors: Eva Michelaraki and George Yannis

Received: 1 December 2025
Revised: 22 January 2026
Accepted: 28 January 2026
Published: 2 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Accurate depth perception is a cornerstone of autonomous vehicle safety, enabling reliable scene understanding, obstacle avoidance, and trajectory planning. While LiDAR and radar systems have traditionally provided high-fidelity depth data, they are often expensive, power-intensive, and suffer from limitations in adverse conditions. Stereo vision offers a scalable, cost-effective alternative capable of producing dense depth maps using calibrated camera pairs [1]. Classical stereo algorithms such as block matching and Semi-Global Matching have been widely used due to their simplicity and low computational cost. However, these methods struggle in textureless regions, reflective surfaces, and under varying illumination conditions, which are frequently encountered in road environments, often producing noisy or incomplete disparity maps [2]. These limitations have given rise to

a shift in focus toward deep learning-based stereo matching models, which offer improved accuracy and robustness. Recent models such as DeepPruner [3], Stereo Anywhere [4], and S^3M-Net [5] have demonstrated strong performance in autonomous driving scenarios. These architectures incorporate attention mechanisms, semantic priors, and multi-scale feature fusion to improve disparity estimation in visually challenging environments. However, real-time deployment of these models on embedded platforms remains an open challenge, particularly when balancing accuracy, speed, and energy efficiency. This gap motivates our work. Among the latest advancements, FoundationStereo [6] introduces a zero-shot stereo matching framework that combines convolutional and transformer-based components. It achieves state-of-the-art performance on benchmarks such as Middlebury and ETH3D, and generalizes across domains without fine-tuning. This paper builds on these advances by optimizing FoundationStereo for real-time inference on embedded platforms and integrating it with YOLO-based object detection to enhance spatial awareness. Unlike prior studies focused on static traffic monitoring, we explore feasibility for dynamic onboard systems, positioning stereo vision as a complementary technology to LiDAR and radar for improved safety. In this paper, we integrate FoundationStereo into a real-time perception pipeline optimized for embedded platforms. The system combines stereo depth estimation with fast object detection using YOLO11 [7], with stereo deformation (in the case of wide field-of-view lenses), rectification and TensorRT acceleration. The proposed approach demonstrates improved depth completeness, robustness, and real-time performance, contributing to intelligent decision systems, multi-sensor integration, smart driving assistance, intelligent transport infrastructure and, in the case of processing the data on an edge server, connected vehicle networks.

2. State of the Art

Stereo depth estimation has undergone a significant transformation over the past decade, evolving from classical computer vision algorithms to models based on deep learning. Modern models leverage deep feature extraction, cost volume regularization, and attention mechanisms to capture global context and fine-grained details. Transformer-based designs extend receptive fields and improve generalization across domains [8]. Additionally, lightweight networks optimized for embedded platforms address the trade-off between accuracy and real-time performance, a critical requirement for autonomous mobility [9,10]. Despite these advances, challenges remain in handling occlusions, thin structures, and achieving domain adaptation without extensive retraining, motivating ongoing work in self-supervised learning and synthetic-to-real transfer [11].

2.1. Deep Learning-Based Stereo Models

Recent advances in deep learning have led to stereo matching networks that outperform classical methods in both accuracy and robustness. These models are increasingly tailored for deployment in, but not limited to, autonomous driving systems and navigation.

- DeepPruner [3] introduced a differentiable PatchMatch module that prunes disparity candidates efficiently, reducing the computational burden of constructing a full cost volume. Instead of evaluating all possible disparities across the stereo image pair, the network learns to select a sparse set of promising candidates for each pixel. This approach enables near real-time inference while maintaining competitive accuracy on KITTI and SceneFlow benchmarks.
- Stereo Anywhere [4] introduced a dual-branch architecture that fuses geometric stereo constraints with monocular depth priors from vision foundation models. It achieves robust zero-shot generalization and handles challenging cases such as mirrors and transparent surfaces by leveraging large-scale monocular priors alongside stereo cues.

The authors also present the MonoTrap dataset, designed to evaluate stereo systems under optical illusions and perspective ambiguities.

- S^3M-Net [5] proposed a joint learning framework that combines semantic segmentation and stereo matching. This integration improves structural consistency and scene understanding, which is critical for autonomous navigation. By introducing modules for feature fusion and semantic consistency-guided loss, S^3M-Net ensures that disparity maps respect object boundaries and semantic structures, reducing errors in occluded or textureless regions.
- FAMNet [12] is a lightweight stereo matching network designed for real-time depth estimation in autonomous driving. It uses attention-guided cost volume construction and multi-scale aggregation to balance accuracy and efficiency. The Fusion Attention-based Cost Volume module reduces reliance on computationally expensive 3D convolutions, while multi-scale attention aggregation enhances disparity prediction through hierarchical feature integration. FAMNet achieves significant improvements over baseline models on KITTI while maintaining low latency, making it suitable for deployment on resource-constrained automotive platforms.
- Cross-Spectral Gated RGB Stereo [13] combines gated imaging with stereo HDR cameras to improve depth estimation in low-light and long-range scenarios, outperforming LiDAR in certain conditions. By fusing RGB, near-infrared, and active illumination cues, this approach addresses certain limitations of passive stereo and single-modality sensors, even in some adverse weather or nighttime environments.

2.2. Transformer-Based Stereo Matching

Transformer-based architectures have recently been applied to stereo depth estimation and have proven to be a compelling alternative to convolutional neural networks (CNNs), offering solutions to several challenges in the field. Traditional CNN-based methods typically rely on local receptive fields, which limits their ability to capture long-range dependencies across stereo image pairs. This constraint becomes especially problematic in scenes with large disparity variations, occlusions where parts of the scene are visible in only one of the stereo views, or textureless regions where accurate pixel correspondence is needed but difficult to achieve. As a result, CNNs often struggle to establish accurate pixel correspondences in such cases, leading to degraded depth estimation performance. Transformers utilize self-attention mechanisms that allow each pixel to compare itself to all others in the image [14]. In stereo matching, this enables the model to reason globally across the left and right views, rather than relying solely on local context. This capability allows Transformer-based models to better overcome the locality constraints of CNNs and better handle complex geometric relationships and occlusions.

- One such example is the STereo TRansformer [15]. STTR formulates stereo matching as a sequence-to-sequence prediction task, leveraging a pure transformer architecture to directly estimate disparities without constructing a traditional cost volume. The model employs cross-attention to establish correspondences between features from the left and right images, and self-attention to enhance contextual understanding within each view. This design enables STTR to reason globally along epipolar lines, making it particularly effective in handling large disparity ranges, occlusions, and textureless regions. STTR uses positional encodings along the horizontal axis, which aligns with the epipolar geometry of rectified stereo pairs. This allows the model to maintain spatial awareness while reducing computational complexity.
- Context-Enhanced Stereo Transformer [16] builds on STTR by introducing the Context Enhanced Path (CEP) module, which captures global scene information beyond epipolar constraints. This addition addresses failure cases in uniform or textureless regions

where local cues are insufficient. By integrating CEP into the Transformer pipeline, CSTR improves performance in zero-shot synthetic-to-real scenarios.

- FoundationStereo [6] is a foundation model for stereo depth estimation. It leverages large-scale synthetic training combined with a self-curation pipeline to ensure data diversity and quality. Architecturally, FoundationStereo integrates Transformer modules with CNN backbones and introduces side-tuning adapters to incorporate monocular priors from vision foundation models. This hybrid approach bridges the sim-to-real gap and enables strong zero-shot generalization across diverse domains, including indoor, outdoor, reflective, and transparent surfaces. Trained on over one million synthetic stereo pairs, it achieves state-of-the-art performance on benchmarks such as Middlebury and ETH3D, and generalizes across domains without fine-tuning.

While this work employs FoundationStereo due to its strong zero-shot generalization, it is important to acknowledge an only recently released model, S^2M^2 [17], as a highly competitive alternative. Unlike FoundationStereo, which combines Transformer modules with CNN backbones and leverages monocular priors, S^2M^2 adopts a fully Transformer-based design without relying on cost volume filtering or deep refinement stacks. Instead, it integrates a multi-resolution Transformer for robust long-range correspondence and introduces a novel loss function that concentrates probability on feasible matches, enabling joint estimation of disparity, occlusion, and confidence. Currently, the model sits at the top of several stereo matching leaderboards, outperforming FoundationStereo in terms of fine detail reconstruction, handling of reflective and transparent surfaces, and even inference speed. Both models share the same goal, which is to improve stereo correspondence, but differ in architectural philosophy. For the scope of this work, the specific choice of model is less critical than the underlying principle: Transformer-based stereo architectures can significantly enhance depth estimation for automotive camera pairs, regardless of whether FoundationStereo, S^2M^2 or any other model is used.

3. Materials and Methods

The proposed perception pipeline integrates stereo depth estimation with object detection to enhance spatial awareness. The pipeline consists of four main stages: image acquisition from stereo camera pairs, fisheye correction (if needed) and rectification, depth estimation using a Transformer-based model, and converting the depth to a point cloud. A schematic of the pipeline is shown in Figure 1. Beyond local inference, the full proposed pipeline architecture described in Section 5 supports an edge-assisted mode. In this configuration, full-resolution stereo pairs, camera intrinsics/extrinsics, and object detection metadata together with the location and speed of the vehicle can be transmitted to an edge server for high-fidelity depth estimation and 3D reconstruction. This would enable advanced use cases such as HD map updates and collaborative environment modeling. First experiments were conducted using wide field-of-view (FOV) fisheye cameras deployed on static traffic monitoring systems. In order to accurately describe the reprojection characteristics of the individual cameras, the parameters of a camera model are optimized based on captured calibration images. Once the stereo system is assembled, a stereo calibration is performed to estimate the relative positions and rotations of the two cameras. These sets of intrinsic and stereo calibration parameters serve as a basis for the rectification of the stereo images. To ensure geometric consistency, fisheye distortion was corrected using OpenCV's distortion model, followed by stereo rectification to align epipolar lines. Rectification involved computing intrinsic parameters (focal length, principal point, distortion coefficients) and extrinsic parameters (rotation and translation between cameras). After rectification, both images shared a unified intrinsic matrix, enabling accurate disparity computation. Invalid regions introduced by rectification were cropped using a custom

algorithm that adjusts the principal point and scales the intrinsics proportionally to the cropped dimensions. Pixels with intensity below a threshold (near black) are identified as invalid. The bounding box of valid pixels is computed and enlarged by a few pixels to eliminate edge artifacts. The intrinsic matrix was updated to reflect the new image coordinate origin defined by the crop window. The principal point (c_x, c_y) is translated by the offset $(crop_x, crop_y)$ of the crop's top-left corner in the original image coordinates, while the focal lengths f_x and f_y remain unchanged in pixel units. For a crop defined by the pixel coordinates of its top-left corner in the original image, the adjusted intrinsic matrix K_{crop} is given by:

$$K_{\text{crop}} = \begin{bmatrix} f_x & s & c_x - crop_x \\ 0 & f_y & c_y - crop_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (1)$$

Here, $(crop_x, crop_y)$ represents the horizontal and vertical pixel offsets defining the crop region's origin. This translation ensures that the principal point correctly represents the projection of the camera's optical center in the new, cropped image coordinate frame. This step was critical because deep learning models interpret black borders as regions of high contrast, leading to severe depth errors as masking these regions out was not a possibility. Cropping ensures geometric consistency and prevents errors in disparity predictions. All preprocessing operations follow well-established formulations from the literature. Camera radial/tangential distortion correction use Zhang's method [18], while fisheye projection adheres to the generic model for wide-angle lenses [19]. Stereo rectification and epipolar alignment are based on multiple-view geometry and rectifying homographies [20,21]. We retain Equation (1) because the cropped-intrinsics update is specific to the configuration and directly affects geometric consistency.

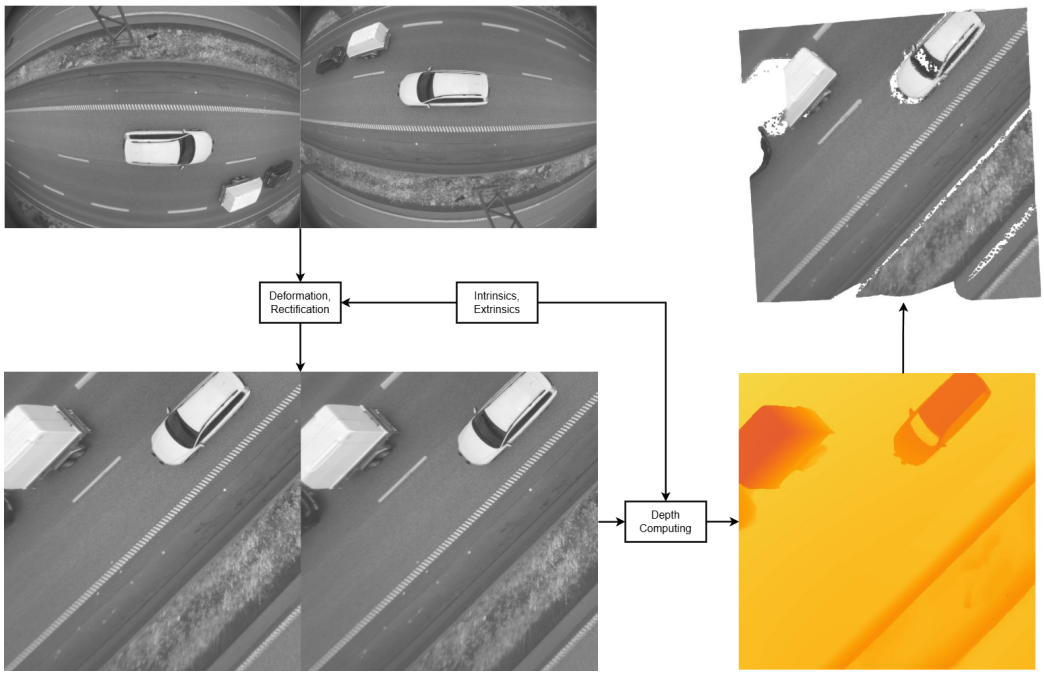


Figure 1. Depth estimation processing pipeline.

Depth estimation was performed using the publicly available FoundationStereo large model, a hybrid architecture combining convolutional layers for local feature extraction with Transformer modules for global context reasoning. The large model was optimized for embedded deployment through TensorRT conversion from PyTorch 2.4 to ONNX 1.16 to TensorRT 8.6, FP16 precision to reduce memory footprint and improve inference speed, and

the power-of-two input resolutions (256 × 256 and 512 × 512) for computational efficiency and alignment with CNN downsampling operations. Using input sizes that are divisible by 2 ensures that these operations can be applied consistently across layers without rounding or cropping, which improves computational efficiency and avoids structural mismatches during inference. Inference was executed first on NVIDIA A6000 and on NVIDIA Jetson AGX Orin platform (64 GB), achieving real-time performance under optimized conditions. Focal lengths f_x and f_y, principal point coordinates c_x and c_y, and skew s all scale inversely with the downsampling factor. Each parameter is divided by the factor by which the image resolution is reduced. This ensures that the geometric relationship between pixels and 3D rays remains consistent, preserving accurate depth estimation despite the change in image size. To validate the geometric correctness of the pipeline and to benchmark the stereo depth estimation component under controlled conditions, the KITTI 2015 [22,23] stereo dataset was used. KITTI provides rectified stereo pairs with precise ground truth disparity, making it suitable for verifying the accuracy of the depth estimation model. The dataset includes per frame calibration files containing intrinsic matrices and the stereo baseline needed, enabling consistent conversion between disparity and metric depth. These calibration parameters were used to confirm that the intrinsic scaling and cropping operations described above preserve the correct projection geometry, serving as a standardized reference dataset to assess the behavior under known imaging conditions. YOLO11 was integrated into the pipeline to provide semantic context for detected objects based on the previous research [24,25]. Detection outputs were fused with depth maps to generate depth-aware bounding boxes, enabling precise localization of detected objects. YOLO inference was also accelerated using TensorRT and executed in parallel with depth estimation. Static Cameras (Traffic Monitoring and Logging) were used for vehicle volume estimation and volume classification in real-world road environments. Preliminary tests also assessed performance under conditions representative of in-vehicle deployment. Datasets included grayscale stereo pairs captured under varying illumination and weather conditions. Approximately 100 image pairs from European and U.S. road sites were processed.

4. Results

FoundationStereo pipeline was tested on NVIDIA A6000 and Jetson AGX Orin using TensorRT FP16 inference. Two input resolutions were evaluated: 256 × 256 and 512 × 512. At 256 × 256, inference only latency was for NVIDIA A6000 was 16.362 per frame (excluding warmup), enabling up to 61.1 FPS for depth estimation alone. At 512 × 512, latency increased to 94.772 ms per frame (excluding warm-up), corresponding to 10.55 FPS, but produced denser point clouds. However, in the case of the static cameras, when using a larger model, the density of the point cloud was the only major difference. End-to-end pipeline latency (pre-processing the image, deformation of the images, rectification, depth computing, YOLO detection and point cloud generation) built in python environment, seen in Figure 2, was approximately 260 ms for 256 × 256, and 367 ms for 512 × 512 resolution. These correspond to 3.85 and 2.75 frames per second, but it should be noted the pipeline was held back by the programming language limitations and by the use of blocking operations. Additionally, AGX Orin combines ARM Cortex-A78AE CPU cores, an Ampere-based GPU, and dedicated accelerators (DLAs, ISP) within a single SoC. This integration minimizes interconnect overhead and enables deterministic scheduling of inference alongside peripheral tasks. By contrast, RTX A6000 operates as a discrete GPU, requiring PCIe communication with a host CPU, which becomes a bottleneck in latency sensitive applications. A re-implementation in C++ using non blocking methods, parallel processing specifically tailored to Jetson AGX Orin is in works, since the ARM cores are optimized for embedded real-time operations, allowing ROS nodes, CAN bus handling, and safety-critical routines

to run in parallel with AI inference. At 256 × 256, inference only latency for Jetson AGX Orin was 38.04 ms per frame (excluding warmup), enabling 26.28 FPS for depth estimation alone. For 512 × 512 resolution the inference time of the neural network was 198.3 ms per frame, enabling 5.04 frames per second, which is unsuitable for real world application. Table 1 summarizes performance metrics for both configurations.

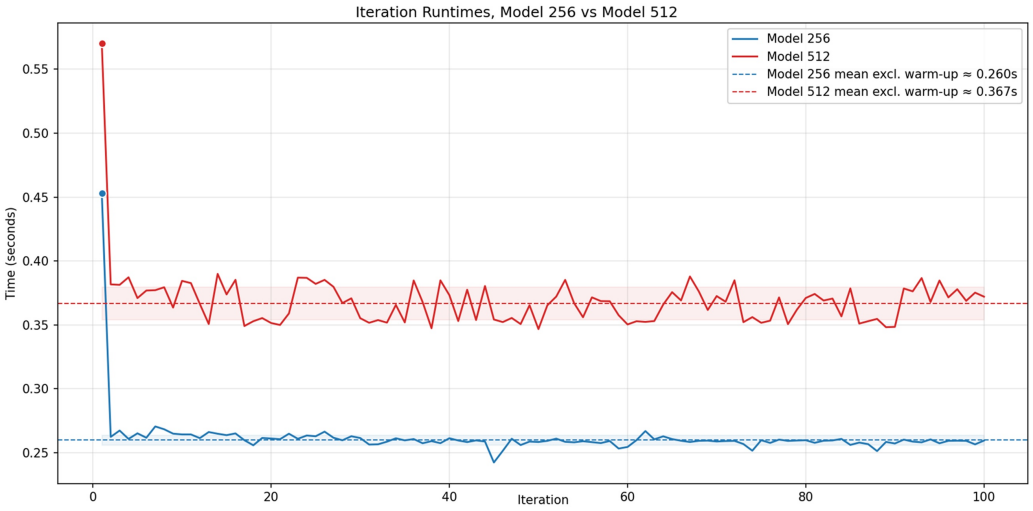


Figure 2. Performance comparison between 256 × 256 (blue) and 512 × 512 (red) model on NVIDIA A6000.

Table 1. Inference latency and FPS comparison for NVIDIA RTX A6000 and Jetson AGX Orin at different input resolutions.

Device	Resolution	Latency (ms)	FPS
RTX A6000	256 × 256	16.362	61.1
RTX A6000	512 × 512	94.772	10.55
AGX Orin	256 × 256	38.04	26.28
AGX Orin	512 × 512	198.3	5.04

4.1. Accuracy Analysis of the Static Cameras

Depth accuracy in the case of static cameras was assessed by comparing reconstructed point clouds to ground truth road meshes using ray-tracing and sampling methods for two separate camera pairs. Given the approximate distance from the camera to the road is $H = 6\,\text{m}$, the recommended baseline-to-height ratio is:

$$B = \alpha \cdot H,\ \alpha \in [0.6, 1.0] \tag{2}$$

Applying this ratio:

$$B = 0.6 \times 6 = 3.6\,\text{m} \text{ to } B = 1.0 \times 6 = 6.0\,\text{m} \tag{3}$$

We used a 3.3 m and 3.6 m base distance between left and right camera. We used two different methods: Sampling and Raytracing. In the Sampling approach, points were uniformly selected from the mesh surface and compared to the point cloud created by 256 × 256 resolution FoundationStereo output, using distance metrics. In contrast, the Raytracing approach casts rays along normals toward the mesh and measures distances to the point cloud. This technique captures surface fidelity more precisely and accounts for occlusions and visibility. Since this was done in the case of the static cameras, the X and Y coordinates were set, the point clouds were already aligned to the Virtual Alignment Point

(VAP), we were comparing the height difference Z. Maximum deviation was 0.044 m in the case of the 3.3 m baseline camera pair. Standard deviation, again, remained higher for a lower baseline pair, approximately 0.01 m, when using the 3.6 m baseline, the results were approximately 0.006 m. Table 2 presents detailed accuracy statistics and Figure 3 provides a visual representation of the deviation histogram.

Table 2. Deviation metrics for Sampling and Raytracing methods at different baselines.

Method	Baseline (m)	Mean (m)	Max (m)	Std (m)
Sampling	3.6	0.0114	0.0369	0.0058
Sampling	3.3	0.0144	0.0444	0.0095
Raytrace	3.6	0.0094	0.0344	0.0065
Raytrace	3.3	0.0123	0.0441	0.0106

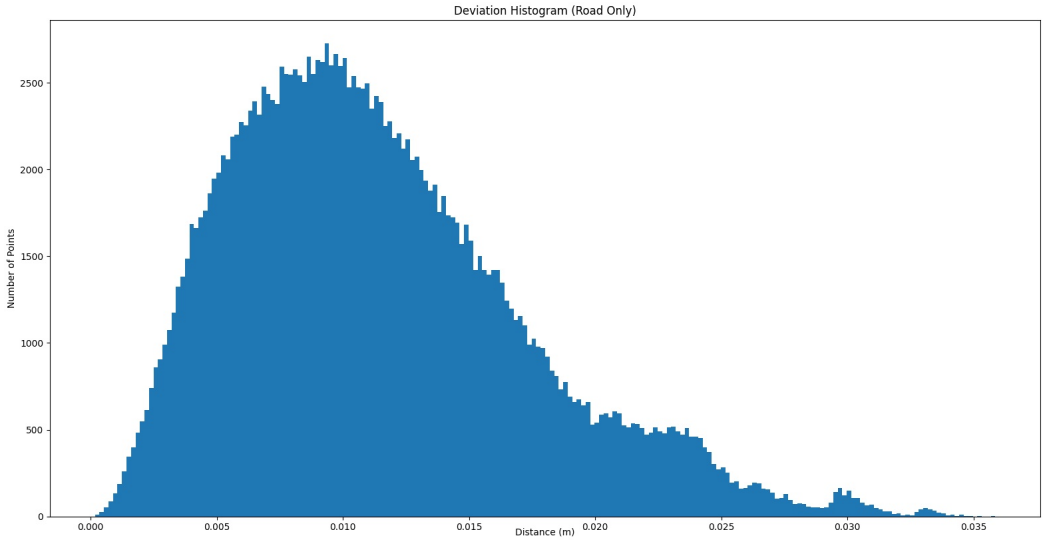


Figure 3. Road surface Z deviation histogram.

4.2. Qualitative Results

To illustrate how baseline distance influences the quality of stereo-based depth estimation in road safety environments, we inspected the reconstructed point clouds and depth maps. In practical road scenarios, reliable perception of object boundaries, lane geometry, and obstacle distances are essential for tasks such as collision avoidance, lane keeping, and vehicle detection. The following examples therefore highlight how different baselines affect the clarity and interpretability of these scene elements. Figure 4 presents point clouds generated at two image resolutions (256 × 256 and 512 × 512) using the 0.3 m baseline. In a road context, higher resolution is expected to preserve sharper edges of vehicles, curbs, and lane markings. This is reflected in the 512 × 512 reconstruction, where object contours appear more coherent and less noisy, in addition to the point cloud being more dense. Such improvements are important for distinguishing between closely spaced objects or identifying road boundaries and play an important role when comparing speed and quality tradeoffs. Figure 5 compares point clouds obtained using the 0.3 m baseline and the longer 3.6 m baseline. A larger baseline generally improves depth determination, as discussed previously in Section 4.1, especially for mid-range objects such as parked cars and roadside infrastructure. This is visible in the 3.6 m reconstruction, where surfaces appear more continuous and depth variation is more pronounced. In contrast, the 0.3 m baseline exhibits a noticeable cutoff and reduced depth range, which could limit the system's ability to reliably estimate distances in real driving situations. This trend was consistent across both the

256 × 256 and 512 × 512 resolutions, indicating that baseline length had a more substantial impact on the quality of the reconstructed geometry than the choice of image resolution.

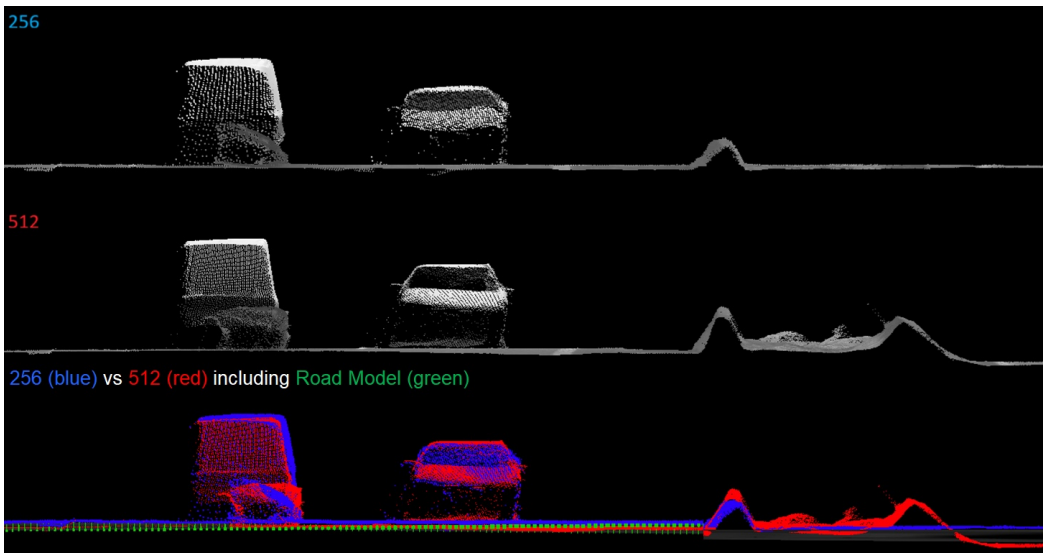


Figure 4. 256 × 256 (blue) vs. 512 × 512 (red) resolutions with 0.3 m baseline including the road model (green).

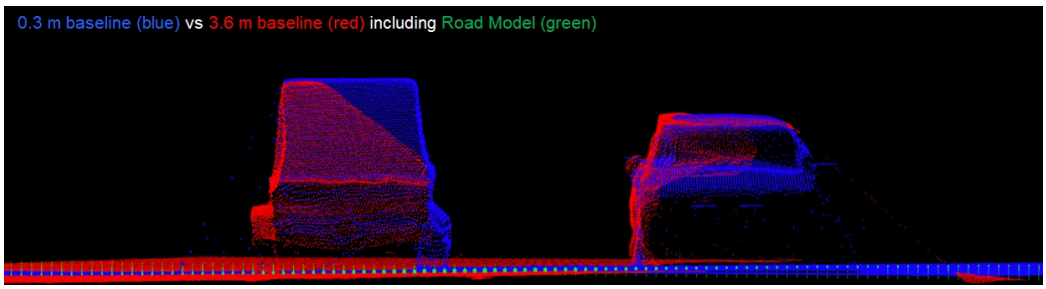


Figure 5. Baseline comparisons in point clouds, 0.3 m baseline (blue) vs. 3.6 m baseline (red) including the road model (green).

Given that the 0.3 m and 3.6 m baselines captured the scene at slightly different angles, the baselines with the most amount of same data were chosen. However, a cutoff is present in the 0.3 m baseline, as can be seen in Figure 6. The present analysis focuses on qualitative inspection, future work will include quantitative comparisons against known object sizes (e.g., standard vehicle dimensions, as is discussed further) to evaluate quantitatively the practical impact of baseline and resolution selection. The pipeline maintained stability under varying illumination and occlusions. Cropping invalid regions significantly improved depth consistency but reduced field of view. This trade-off was mitigated by multi-camera tiling in static deployments, a common practice even in previously used depth estimation methods, as seen in Figure 7, where it is also compared to the results of a traditional approach used in volume computing. This lightweight method operates on processed edge maps rather than raw grayscale images, scanning epipolar lines to match edge features between stereo pairs. The resulting disparity values are converted into depth using standard stereo geometry. While this approach produces sparse point clouds concentrated along structural contours, its simplicity and low computational overhead enable real-time performance, making it highly effective for embedded devices in scenarios where speed is critical. The resulting tiled Transformer point cloud is much more dense, even at 256 × 256 resolution, capturing the details and shape more precisely than the previous methods.

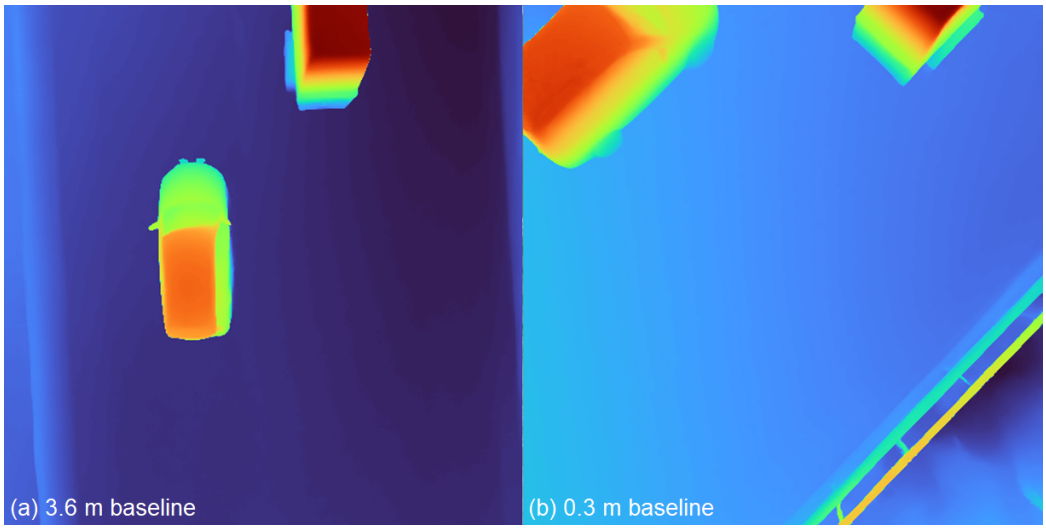


Figure 6. Comparison of depth estimation for two different baselines. (**a**) for 3.6 m baseline, and (**b**) for 0.3 m baseline.

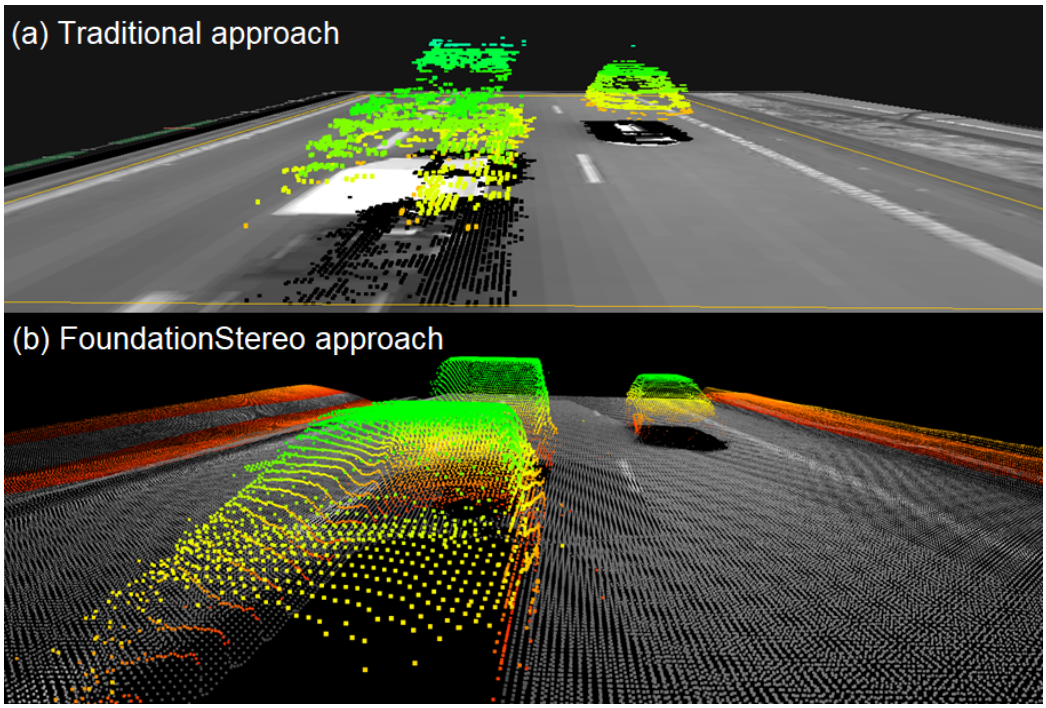


Figure 7. Comparison of point clouds generated by the (**a**) traditional method and (**b**) FoundationStereo Transformer-based approach.

4.3. Textureless Regions, Reflective Surfaces and Varying Illumination Conditions

FoundationStereo demonstrates strong robustness in scenarios with varying illumination and reflective surfaces, primarily because its learned representations are less dependent on strict photometric consistency and can generalize across domains. This makes it particularly effective in handling specular highlights and lighting changes compared to traditional stereo pipelines. Despite the presence of inaccuracies in object shape representation, for example in the shape of the wheels of the vehicle, particularly during nighttime with illumination on reflective surfaces, significant number of textureless regions and an incomplete vehicle shape captured on frame, the overall 3D positioning of XYZ coordinates of the vehicle demonstrated stability across frames, particularly when using two camera pairs and using previously mentioned tiling, as can be seen in Figure 8. This consistency enables

reliable volume estimation and spatial reasoning, highlighting the model's capacity to maintain geometric integrity despite photometric variations.

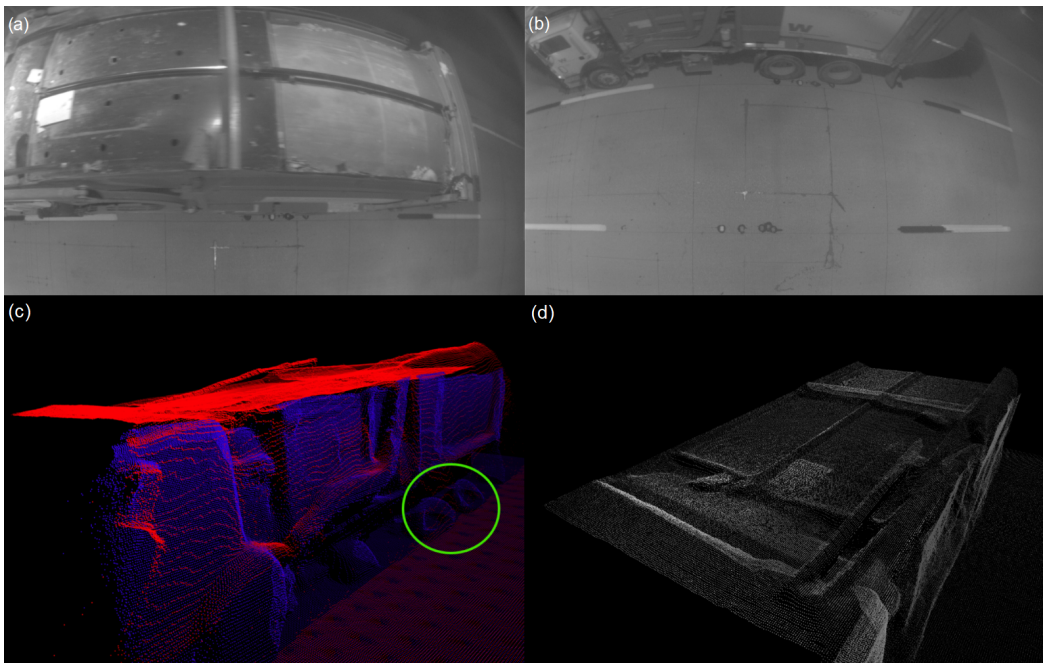


Figure 8. The nighttime scene contains textureless regions, reflective surfaces, and glare from nighttime illumination. (**a**,**b**) represent the input images, (**c**) shows consistent dimensions with mistakes highlighted in green, and (**d**) shows the detail and precision of the reconstructed roof of the vehicle.

We observed that the model's resolution plays a role in handling reflective surfaces. Counterintuitively, the 256 × 256 resolution variant demonstrated superior performance in correctly interpreting reflective surfaces compared to the 512 × 512 resolution version as can be seen in Figure 9, likely due to better regularization against overfitting to photometric artifacts that become more pronounced at higher resolutions. We note that this observation is based on specific examples and may not constitute a general statistical trend, warranting further investigation to validate this effect across a broader area.

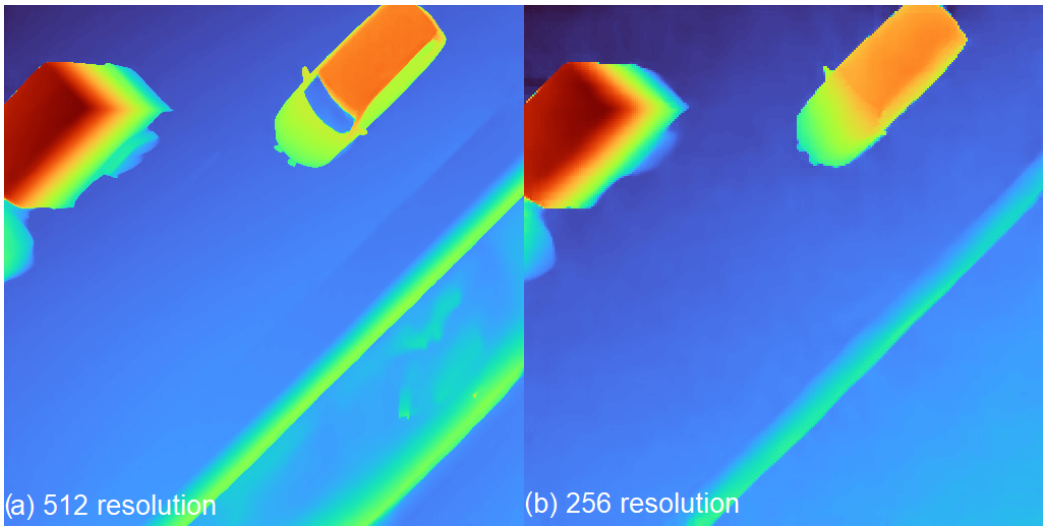


Figure 9. (**a**) 512 × 512 vs. (**b**) 256 × 256 resolutions of FounDationStereo handling reflective surfaces.

4.4. Roadside Conditions

To evaluate the proposed pipeline under realistic roadside conditions, we captured stereo images using a dual camera setup with a baseline of 0.15 m on a road. Initially, low-resolution stereo pairs were processed with FoundationStereo to provide real-time depth estimation for vehicle navigation and situational awareness. That being said, the base distance is not ideal. For high-resolution reconstruction, rectified Full HD stereo images (1230 × 755 px) were processed through the FoundationStereo model to generate dense depth maps of the environment. Before converting these depth maps into point clouds, we applied YOLO11 object segmentation on the left image to segment vehicles and other dynamic objects present on the road. The detected object coordinates were then used to mask corresponding regions in the depth map, effectively removing unwanted elements from the scene. Finally, the filtered depth maps were reconstructed into 3D point clouds, resulting in a clean and detailed scan of the roadside environment, as can be seen in Figure 10. This approach demonstrates how combining advanced stereo depth estimation with object segmentation enables accurate and clutter-free environmental modeling.

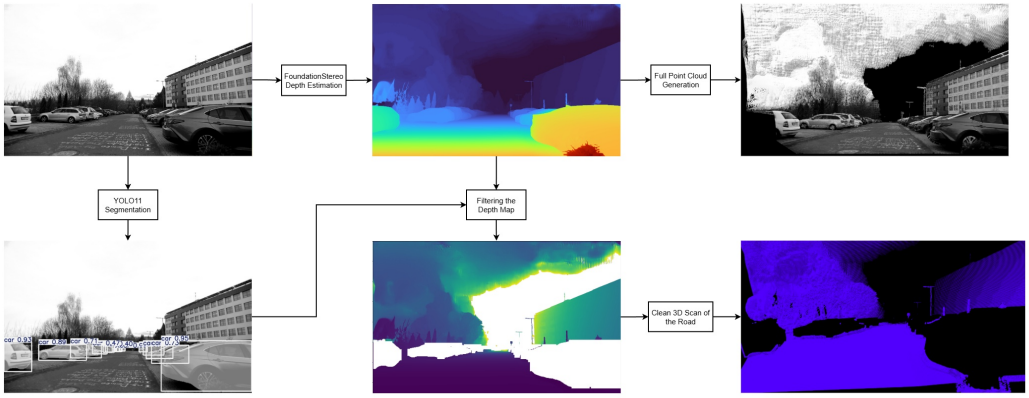


Figure 10. Camera to point cloud pipeline, with depth filtering based on object detection using YOLO11 segmentation.

4.5. KITTI 2015 Stereo Accuracy Evaluation

The KITTI evaluation in this chapter focuses on establishing a controlled, reproducible procedure for running FoundationStereo at a resolution compatible with the embedded deployment constraints. All KITTI stereo pairs were processed with the described pipeline. To match the computational budget of the 256 × 256 embedded configuration, each stereo pair was uniformly downscaled from 1242 × 375 so that the image height became 160 pixels. A centered horizontal crop reduced the width to 448 pixels, removing 41 pixels from each side. This yields a 448 × 160 input that is divisible by the model's architecture and preserves most of the vertical field of view. The intrinsic matrix was scaled proportionally to the image downsampling factor. Focal lengths and principal point coordinates scaled as described in Section 3. Because the crop removes pixels from the left and right edges, the intrinsic matrix is updated by subtracting from c_x while leaving c_y unchanged, as no vertical cropping occurs. After depth is predicted, it is upsampled back to the original KITTI height using the inverse scaling factor. This restores the full vertical resolution of 375 pixels while preserving the reduced horizontal field of view introduced by the pixel crop. Because the crop removes 82 pixels in the scaled domain, the upsampled result corresponds to a central strip of 1050 × 375 pixels in the original KITTI coordinate frame. The missing left and right regions are not reconstructed, as the crop defines the effective field of view used during inference. To ensure a valid comparison with KITTI ground truth, the corresponding central pixel region of the ground truth disparity map was extracted using the same crop geometry. All disparity and depth-based metrics were then computed only on this overlapping region,

ensuring that evaluation is performed on a geometrically consistent field of view that matches the model's effective input. The accuracy was quantified using four widely used metrics that capture both pixel-level disparity errors and their effect on metric depth reconstruction. All metrics were computed on the valid region of the cropped KITTI field of view, defined by positive ground truth disparity and finite predicted depth. End-Point Error (EPE) measures the mean absolute disparity error in pixels. D1 Outlier Rate reports the percentage of pixels whose disparity error exceeds the KITTI threshold of 3 px or 5% of the ground truth disparity. Absolute Relative Error (AbsRel) quantifies the proportional depth error by normalizing the absolute depth difference by the ground truth depth. Root Mean Square Error (RMSE) expresses the magnitude of depth deviations in meters. The results in Table 3 show that FoundationStereo maintains competitive accuracy under resolution reduction imposed by the embedded deployment constraints. At an effective input of 448 × 160 pixels, the model achieves an average EPE of 0.901 px and a D1 outlier rate of 3.483%, indicating that most disparity estimates remain within a small pixel error. The depth-based metrics follow the same trend: an AbsRel of 0.075 and an RMSE of 20.387 m reflect stable metric depth reconstruction across the evaluated scenes. The narrow 95% bootstrap confidence intervals further confirm that the performance is consistent across the full KITTI 2015 set.

Table 3. KITTI 2015 stereo accuracy at an input resolution of 448 × 160 pixels.

Resolution	EPE (px)	D1 (%)	AbsRel	RMSE (m)
448 × 160	0.901	3.483	0.075	20.387

Note: 95% bootstrap confidence intervals are: EPE [0.868, 0.937], D1 [3.208, 3.788]%, AbsRel [0.0611, 0.0957], RMSE [8.43, 41.79] m.

To complement the aggregated KITTI statistics, Figure 11 provides a breakdown of a representative frame. Frame 000079_10 was selected as a representative example because its error statistics lie close to the dataset-level mean. This visualization illustrates how the evaluation pipeline behaves on a per image basis and highlights the spatial structure of the model's disparity errors. Panel (a) presents the absolute disparity error map, computed as the pixel-wise difference between the predicted and ground truth disparity. To preserve interpretability across the full dynamic range, the colormap is clipped at the 99th percentile of the error distribution, ensuring that subtle variations in low error regions remain visible. Panel (b) visualizes the D1 outlier mask, where pixels exceeding the KITTI threshold of 3 px or 5% of the ground truth disparity are highlighted in cyan. This mask isolates the regions where the model fails to meet the KITTI accuracy criteria.

Beyond the visual analysis, the quantitative results for this frame further confirm its representativeness. As can be seen in Table 4, the model attains a D1 outlier rate of 2.800%, aligned with the dataset-level averages reported in Table 3. The depth-based metrics follow the same trend. An AbsRel of 0.101 and an RMSE of 20.417 m indicate that the metric depth reconstruction for this frame is consistent with the overall behavior observed across the full KITTI 2015 evaluation. The RMSE99 disparity error of 1.096 px further shows that, after removing extreme outliers, the residual disparity errors remain low. Overall, while these values are higher than those reported by state-of-the-art stereo networks evaluated at the native 1242 × 375 KITTI resolution (typically around 1–2% D1 on the official benchmark), they remain competitive given the substantial reduction in input size and effective field of view. In particular, the elevated D1 rate should be interpreted in the context of the evaluation setup: the downscaling removes fine-grained geometric detail and the KITTI D1 threshold becomes disproportionately strict at low resolution. The depth-based metrics further indicate that the model maintains stable metric depth

reconstruction under the imposed embedded scale constraints, demonstrating that the observed errors arise primarily from the challenging reduced resolution regime.

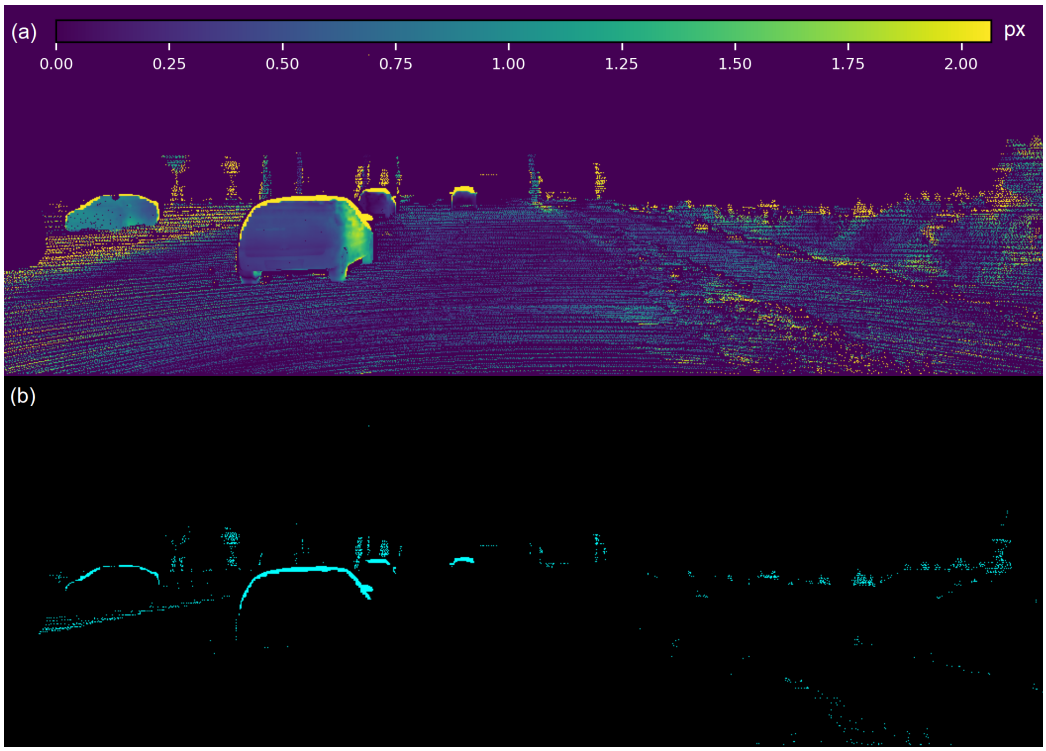


Figure 11. Evaluation on KITTI 2015 for frame 000079_10. (**a**) Absolute disparity error map with 99th-percentile clipping in pixels. (**b**) D1 outlier mask highlighting pixels whose disparity error exceeds the KITTI threshold.

Table 4. Single–frame KITTI 2015 accuracy for representative frame 000079_10.

Frame	EPE (px)	D1 (%)	AbsRel	RMSE (m)
000079_10	0.896	2.800	0.101	20.417

5. Discussion

The results demonstrate that Transformer-based stereo depth estimation, when optimized for embedded platforms, can achieve real-time performance while maintaining high levels of accuracy. This capability has significant implications for both static and dynamic deployment scenarios. Beyond local inference, the proposed pipeline seen in Figure 12 enables cooperative perception through edge computing. Vehicles transmit full-resolution stereo pairs, camera parameters, detection boxes, and metadata (e.g., GPS, speed) to an edge server, where high-resolution models reconstruct detailed 3D scans of the environment. To focus on infrastructure mapping, dynamic objects such as vehicles and pedestrians are removed from depth maps using object detection or segmentation (e.g., YOLO11) prior to point cloud generation. This filtering produces clean road geometry and reduces post-processing complexity compared to LiDAR-based workflows, where object removal occurs at the point cloud level. Although gaps remain where objects were excluded, these can be progressively filled as additional vehicles contribute data. Aggregating multiple filtered point clouds supports progressive HD map updates, consistent with 5GAA and 3GPP V2X standards [26,27]. For high-quality reconstructions, the edge server applies high-resolution segmentation and stereo networks, justifying the transmission of full-resolution imagery. While stereo vision is generally still less precise than LiDAR for depth estimation, literature and praxis indicates it is a viable and cost-effective alternative when LiDAR is unavailable

or unusable, offering dense depth maps and scalability through existing camera infrastructure [28–30]. This approach offers a scalable alternative to traditional HD mapping by leveraging existing camera infrastructure and connectivity.

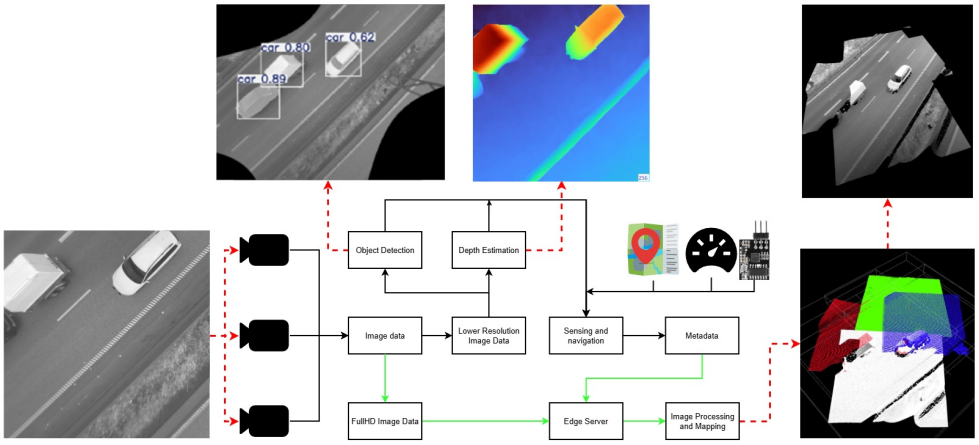


Figure 12. Proposed image processing pipeline. Red lines indicate visual input/output, black lines indicate internal onboard processes, green lines indicate outgoing data.

While the pipeline achieves real-time performance at reduced resolutions, full-HD inference remains computationally intensive. Cropping invalid regions improves geometric consistency but reduces field of view, which may limit situational awareness. Bandwidth constraints [31,32] and privacy concerns must be addressed when transmitting images to edge servers. Furthermore, stitching point clouds from multiple vehicles requires precise synchronization and confidence mapping to mitigate dead zones.

6. Conclusions

This work demonstrates the capabilities of Transformer-based stereo depth estimation, improving perception for autonomous mobility. By integrating FoundationStereo with YOLO11, the proposed pipeline provides both geometric and semantic understanding of the environment, enabling safer navigation and more reliable decision-making. In traffic monitoring applications, among other uses, depth maps generated from static cameras can be converted into point clouds to estimate vehicle volume. This enables more accurate classification of vehicle types. Such functionality supports enforcement, tolling, and traffic analytics, contributing to improved road safety and infrastructure planning. While the results are encouraging, further investigation is warranted. Future work would involve using a vehicle with known dimensions to validate the estimated volume and dimensions against ground truth measurements. For onboard systems, enhanced depth maps improve spatial awareness, enabling safer navigation and collision avoidance. By complementing LiDAR and radar, this approach to stereo vision provides dense depth information. This integration strengthens perception in scenarios where some active sensors may fail, such as adverse weather or reflective surfaces. Furthermore, leveraging such methods enhances depth perception without replacing existing sensors, contributing to sensor fusion. When all modalities deliver high-quality data, the decision-making process of the onboard computer becomes more robust, ultimately improving navigation reliability.

Beyond local inference, this paper introduces a vision for edge-assisted HD map reconstruction. Vehicles can upload full-resolution stereo pairs, intrinsics/extrinsics, detection boxes, and metadata (e.g., GPS, speed) to an edge server, where high-resolution models reconstruct detailed 3D scans. By filtering detected objects before point cloud generation,

the system produces clean road geometry, reducing post-processing overhead compared to LiDAR-based mapping. Aggregating multiple filtered point clouds from different vehicles enables progressive HD map updates, aligning with 5GAA and 3GPP V2X standards. This approach offers a scalable alternative to traditional HD mapping, leveraging existing camera infrastructure and edge connectivity.

While the pipeline achieves real-time performance at reduced resolutions, several practical limitations remain that need to be adressed or mitigated. Full-HD inference is still too computationally intensive for embedded platforms; camera constraints (limited baseline, sensor noise, rolling shutter and synchronization errors, fisheye distortion and cropping) reduce effective field of view and can introduce geometric artifacts and latency arises from CPU bound pre/post-processing and blocking I/O. Precision of the network is notable, it ranks highly on the Middlebury and ETH3D leaderboards [33,34] and has already been trained and tried on established datasets like SceneFlow, Sintel, Virtual KITTI 2, FallingThings, TartanAir and CREStereo [6]. Our goal in this work was to evaluate whether FoundationStereo can be applied to road scenarios like the ones studied here (both static roadside monitoring and dynamic, vehicle-mounted use). Stereo is a passive, image-based modality that yields dense relative depth dependent on baseline, texture, illumination, and calibration, it cannot fully replace LiDAR's absolute, long-range, and well characterized range measurements. This paper advocates for sensor fusion rather than replacement, each with their own positives and drawbacks. A direct comparison between LiDAR point clouds and FoundationStereo reconstructions is currently in works. Future work will also focus on GPU-based post-processing to eliminate CPU bottlenecks on the ARM infrastructure, adaptive resolution strategies for balancing accuracy and speed, compression techniques for efficient edge upload. Looking ahead, the forthcoming Fast-FoundationStereo [35] architecture suggests that higher resolution real-time inference may become feasible without the aggressive downscaling required in this study, enabling future evaluations under less restrictive conditions. Integration with V2X protocols for cooperative perception and HD map generation and multi-vehicle point cloud stitching with confidence mapping to mitigate dead zones. By combining local real-time perception with cooperative edge-based reconstruction, this approach lays the foundation for scalable, camera-driven perception systems that enhance road safety and accelerate the transition toward fully connected and automated mobility.

Author Contributions: Conceptualization, M.S., M.D. and A.K.; methodology, M.S., M.D. and A.K.; software, M.S. and A.K.; validation, M.S.; formal analysis, M.S.; investigation, M.S.; resources, M.S. and A.K.; data curation, M.S. and A.K.; writing—original draft preparation, M.S.; writing—review and editing, M.S., M.D. and A.K.; visualization, M.S.; supervision, M.D. and A.K.; project administration, M.D.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Univesity of Žilina Grant System, under the designation "Development of robust location-navigation systems for communication-enabled autonomous mobility in harsh environments" (ID 20407) and by the Slovak Scientific Grant Agency (VEGA) Grant Agency through the Research of Effective Localization and Communication Solutions in Heterogenous Wireless Networks 1/0580/25 (ID 20345).

Data Availability Statement: All used neural networks and datasets are publicly available, and Yolo11 is available at https://docs.ultralytics.com/models/yolo11/, accessed on 20 December 2025. FoundationStereo is available at https://github.com/NVlabs/FoundationStereo, accessed on 20 December 2025. KITTI Scene Flow Evaluation 2015 is available at https://www.cvlibs.net/datasets/kitti/eval_scene_flow.php, accessed on 11 January 2026. Concerning the location of the static camera gantries, located at highway E4 in Jönköping, Sweden. Concerning the precise parameters of the static cameras, this data is not publicly available, since it is the property of Kapsch TrafficCom AG. The images used

are not publicly available due to ethical and privacy considerations, as they contain identifiable license plates and contextual data that qualify as personal data under the GDPR; only data rendered truly anonymous falls outside the regulation, which is not feasible without compromising the integrity of the dataset. Public disclosure would also violate the principles of data minimization and privacy by design. The European Data Protection Board guidance on video devices restricts disclosure of footage without appropriate safeguards.

Acknowledgments: The authors would like to thank Kapsch TrafficCom AG, the Department of Video and Image Analytics, and the Faculty of Electrical Engineering and Information Technology at the University of Žilina for supporting this research.

Conflicts of Interest: Author Andrei Kurdiumov was employed by the company Kapsch TrafficCom AG. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AbsRel	Absolute Relative Error
CEP	Context Enhanced Path
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
c_x, c_y	Principal point coordinates
DLA	Deep Learning Accelerator
EPE	End-Point Error
FOV	Field of View
FPS	Frames Per Second
FP16	16-bit Floating Point
f_x,f_y	Focal lengths
GPU	Graphics Processing Unit
HD	High Definition
HDR	High Dynamic Range
ISP	Image Signal Processor
KITTI	Karlsruhe Institute of Technology and Toyota Technological Institute dataset
LiDAR	Light Detection and Ranging
ONNX	Open Neural Network Exchange
PCIe	Peripheral Component Interconnect Express
RGB	Red Green Blue
RMSE	Root Mean Square Error
ROS	Robot Operating System
S2M2	Scalable Stereo Matching Model
S3M-Net	Semantic Segmentation and Stereo Matching Network
SoC	System on Chip
STTR	STereo TRansformer
VAP	Virtual Alignment Point
V2X	Vehicle-to-Everything
YOLO11	You Only Look Once version 11

References

1. Fan, R.; Wang, L.; Bocus, M.J.; Pitas, I. Computer Stereo Vision for Autonomous Driving. *arXiv* **2020**, arXiv:2012.03194. [CrossRef]
2. Fan, R.; Guo, S.; Bocus, M.J. (Eds.) *Autonomous Driving Perception: Fundamentals and Applications*; Advances in Computer Vision and Pattern Recognition; Springer Nature: Singapore, 2023. [CrossRef]

3. Blankenberg, E.; Blankenberg, S. Survey of Disparity Map Algorithms Intended for Real-Time Stereoscopic Depth Estimation. 2021. Available online: https://www.semanticscholar.org/paper/fe59e7f76602b1979ecd120b0a88fe97b78e1d96 (accessed on 4 November 2025).
4. Bartolomei, L.; Tosi, F.; Poggi, M.; Mattoccia, S. Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail. *arXiv* **2025**, arXiv:2412.04472. [CrossRef]
5. Wu, Z.; Feng, Y.; Liu, C.-W.; Yu, F.; Chen, Q.; Fan, R. S^3M-Net: Joint Learning of Semantic Segmentation and Stereo Matching for Autonomous Driving. *IEEE Trans. Intell. Veh.* **2024**, *9*, 3940–3951. [CrossRef]
6. Wen, B.; Trepte, M.; Aribido, J.; Kautz, J.; Gallo, O.; Birchfield, S. FoundationStereo: Zero-Shot Stereo Matching. *arXiv* **2025**, arXiv:2501.09898. [CrossRef]
7. Ultralytics. YOLO11. Available online: https://docs.ultralytics.com/models/yolo11 (accessed on 4 November 2025).
8. Tosi, F.; Bartolomei, L.; Poggi, M. A Survey on Deep Stereo Matching in the Twenties. *Int. J. Comput. Vis.* **2025**, *133*, 4245–4276. [CrossRef]
9. Chang, J.R.; Chang, P.C.; Chen, Y.S. Attention-Aware Feature Aggregation for Real-Time Stereo Matching on Edge Devices. In *Computer Vision—ACCV 2020*; Ishikawa, H., Liu, C.L., Pajdla, T., Shi, J., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2021; Volume 12622. [CrossRef]
10. Huang, Q.; Zhang, Y.; Zheng, J.; Shang, G.; Chen, G. A CNN-Based Real-Time Dense Stereo SLAM System on Embedded FPGA. In *Artificial Intelligence. CICAI 2023*; Fang, L., Pei, J., Zhai, G., Wang, R., Eds.; Lecture Notes in Computer Science; Springer: Singapore, 2024; Volume 14474. [CrossRef]
11. Iqbal, W.; Paffenholz, J.A.; Mehltretter, M. Guiding Deep Learning with Expert Knowledge for Dense Stereo Matching. *PFG–J. Photogramm. Remote Sens. Geoinf. Sci.* **2023**, *91*, 365–380. [CrossRef]
12. Zhang, J.; Tong, Q.; Yan, N.; Liu, X. FAMNet: A Lightweight Stereo Matching Network for Real-Time Depth Estimation in Autonomous Driving. *Symmetry* **2025**, *17*, 1214. [CrossRef]
13. Brucker, S.; Walz, S.; Bijelic, M.; Heide, F. Cross-Spectral Gated-RGB Stereo Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Seattle, WA, USA, 2024; pp. 21654–21665. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762. [CrossRef] [PubMed]
15. Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F.X.; Taylor, R.H.; Unberath, M. Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers. *arXiv* **2021**, arXiv:2011.02910. [CrossRef]
16. Guo, W.; Li, Z.; Yang, Y.; Wang, Z.; Taylor, R.H.; Unberath, M.; Yuille, A.; Li, Y. Context-Enhanced Stereo Transformer. *arXiv* **2022**, arXiv:2210.11719. [CrossRef]
17. Min, J.; Jeon, Y.; Kim, J.; Choi, M. S2M2: Scalable Stereo Matching Model for Reliable Depth Estimation. *arXiv* **2025**, arXiv:2507.13229. [CrossRef]
18. Zhang, Z. A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
19. Kannala, J.; Brandt, S.S. A Generic Camera Model and Calibration Method for Conventional, Wide-Angle, and Fish-Eye Lenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1335–1340. [CrossRef] [PubMed]
20. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
21. Hartley, R.I. Theory and Practice of Projective Rectification. *Int. J. Comput. Vis.* **1999**, *35*, 115–127. [CrossRef]
22. Menze, M.; Heipke, C.; Geiger, A. Object Scene Flow. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 60–76. [CrossRef]
23. Menze, M.; Heipke, C.; Geiger, A. Joint 3D Estimation of Vehicles and Scene Flow. In *Proceedings of the ISPRS Workshop on Image Sequence Analysis (ISA)*; International Society for Photogrammetry and Remote Sensing (ISPRS): Munich, Germany, 2015. [CrossRef]
24. Jegham, N.; Koh, C.Y.; Abdelatti, M.; Hendawi, A. Evaluating the Evolution of YOLO Models: A Comprehensive Benchmark Study of YOLO11 and Its Predecessors. *arXiv* **2025**, arXiv:2411.00201v1. Available online: https://arxiv.org/html/2411.00201v1 (accessed on 21 November 2025).
25. Simeonov, M.; Kamencay, P.; Dado, M. Evaluating YOLOv11's Role in Robust Real-Time Object Detection for Autonomous Driving. In *Proceedings of RADIOELEKTRONIKA 2025*; IEEE: Bratislava, Slovakia, 2025; p. 5. [CrossRef]
26. 5GAA. C-V2X Roadmap White Paper III. 5G Automotive Association, January 2025. Available online: https://5gaa.org/content/uploads/2025/01/5gaa-wi-cv2xrm-iii-roadmap-white-paper.pdf (accessed on 21 November 2025).
27. ETSI. Intelligent Transport Systems (ITS); Service Requirements for V2X Services. ETSI TS 122 186 V17.0.0 (2025-01). Available online: https://www.etsi.org/deliver/etsi_ts/122100_122199/122186/17.00.00_60/ts_122186v170000p.pdf (accessed on 21 November 2025).

28. IEEE Spectrum. Camera Crushes Lidar, Claims Startup. IEEE Spectrum 2021. Available online: https://spectrum.ieee.org/camera-crushes-lidar (accessed on 26 November 2025).
29. NODAR. Stereo Vision Technology for Autonomous Vehicles. Available online: https://www.nodarsensor.com/ (accessed on 26 November 2025).
30. You, Y.; Wang, Y.; Chao, W.-L.; Garg, D.; Pleiss, G.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR++: Accurate Depth for 3D Object Detection. *arXiv* **2019**, arXiv:1906.06310. [CrossRef]
31. Cucor, B.; Petrov, T.; Kamencay, P.; Pourhashem, G.; Dado, M. Physical and Digital Infrastructure Readiness Index for Connected and Automated Vehicles. *Sensors* **2022**, *22*, 7315. [CrossRef] [PubMed]
32. Cucor, B.; Petrov, T.; Kamencay, P.; Simeonov, M.; Dado, M. Digital Infrastructure Quality Assessment System Methodology for Connected and Automated Vehicles. *Electronics* **2023**, *12*, 3886. [CrossRef]
33. ETH3D. Low-Resolution Two-View Stereo Benchmark. Available online: https://www.eth3d.net/low_res_two_view (accessed on 20 December 2025).
34. Middlebury Stereo Vision. Stereo Evaluation (Eval3). Available online: https://vision.middlebury.edu/stereo/eval3/ (accessed on 20 December 2025).
35. Wen, B.; Dewan, S.; Birchfield, S. Fast-FoundationStereo: Real-Time Zero-Shot Stereo Matching. *arXiv* **2025**, arXiv:2512.11130. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.