IEEE Open Journal of Vehicular Technology

Received 11 March 2026; accepted 15 March 2026. Date of publication 23 March 2026;
date of current version 7 April 2026. The review of this article was coordinated by Editor Itzik Klein.

Digital Object Identifier 10.1109/OJVT.2026.3676610

Monocular Ground Normal Prediction for the Road Ahead

NORBERT MARKÓ [1,2], ZOLTÁN RÓZSA [1,3] (Member, IEEE), ÁRON BALLAGI [2], AND TAMÁS SZIRÁNYI [1,3] (Life Senior Member, IEEE)

[1]Machine Perception Research Laboratory, HUN-REN SZTAKI, 1111 Budapest, Hungary
[2]Vehicle Industry Research Center, Széchenyi István University, 9026 Győr, Hungary
[3]Department of Material Handling and Logistics Systems, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, 1111 Budapest, Hungary

CORRESPONDING AUTHOR: NORBERT MARKÓ (e-mail: marko.norbert@ga.sze.hu).

This work was supported in part by the National Research, Development and Innovation Office and the National Defense Cooperative Doctoral Program 2021 under Grant NVKDP-2021, in part by the European Union within the Framework of the National Laboratory for Autonomous Systems under Grant RRF-2.3.1-21-2022-00002, in part by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund under Grant STARTING 149552 and Grant K139485 through STARTING_24 and K_21 NKFIH funding schemes, and in part by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (MTA).

The proposed pipeline and dataset are publicly available at: https://norbertmarko.github.io/imu_cam_normal_prediction/.
This article has supplementary downloadable material available at https://doi.org/10.1109/OJVT.2026.3676610, provided by the authors.

ABSTRACT Robust fusion of monocular and inertial data has the potential to offer a low-cost alternative for ground surface normal prediction ahead, compared to more expensive sensors, such as LiDAR. Yet robust camera-based prediction remains challenging, particularly for steep grades and texture-poor, homogeneous road surfaces. To address these issues, we propose an enhanced monocular camera-IMU fusion pipeline incorporating a lightweight transformer-based feature matcher for improved correspondence accuracy, and robust temporal filtering, using a spherical linear interpolation (SLERP) filter, to enhance consistency and reduce drift. To enable rigorous benchmarking and reproducibility, we also standardize the evaluation protocol and release a novel dataset containing synchronized camera, LiDAR, and IMU-derived pose data, specifically captured across diverse incline and decline scenarios. Extensive continuous validation demonstrates that our method significantly improves both accuracy and temporal stability over existing approaches, setting a new state of the art for robust, continuous ground normal estimation ahead.

INDEX TERMS Ground normal prediction, IMU-camera fusion, image-based ground plane prediction, transformer, SLERP, road surface normal prediction, road surface pitch prediction.

I. INTRODUCTION

In the push towards safer, more affordable driver-assistance and autonomous systems, ubiquitous, low-cost monocular cameras, paired with miniature inertial measurement units (IMUs), can deliver rich situational awareness at a fraction of the cost of LiDAR or multi-camera setups [1], [2]. Yet, forecasting the road surface's 3D orientation from a single moving camera, especially on uphill and downhill grades, remains a formidable challenge. Predicting this ground normal vector is one of the most important building blocks for reconstructing a 3D scene from monocular images [3]. It enables inverse perspective mapping [4], [5], [6], [7] and planar parallax estimation [8], which in turn supports robust free-space estimation [9]. Higher-level perception modules, such as camera pose estimation [10], 3D object detection [11] and semantic segmentation [12] can also operate with a higher accuracy and reliability. Leveraging the inherent simplicity of late-stage fusion between a monocular camera and an IMU, this work aims to recover the road surface normal with high accuracy and a particular emphasis on robust performance across varying inclines and declines.

Monocular ground normal estimation methods fall into two broad categories. Direct geometric approaches typically utilize feature matching to find planar image correspondences and to extract camera motion and plane orientation from them [13]. Learning-based methods, on the other hand, train

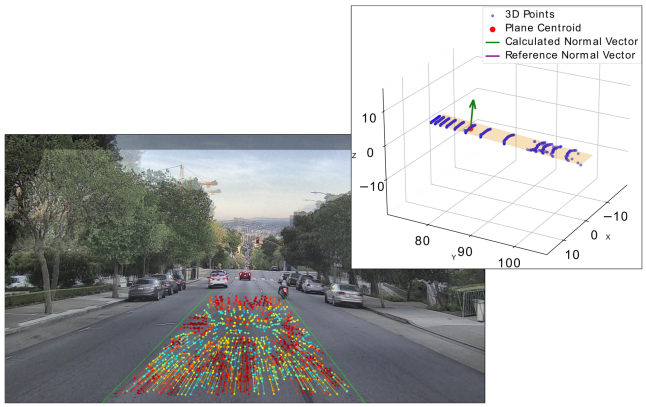

FIGURE 1. Our algorithm's example output on a decline surface. The overlaid point pairs are correspondences produced by the feature matching algorithm, color-coded by matching confidence: colder colors indicate lower confidence and warmer colors indicate higher confidence. A more detailed view of the algorithm's output can be seen on Fig. 6.

neural networks to regress depth and surface normal vectors directly from image patches, as exemplified by GroundNet [14].

Despite advances in the field, three key gaps remain. First challenge is slope sensitivity. Most existing pipelines assume locally straight and flat roads, causing sudden performance degradation on sharp turns and on sustained grades where road orientation and scale become entangled. Second gap is road texture sensitivity. Methods relying on rich road features (e.g., lane markings or asphalt patterns) often fail on homogeneous road surfaces, which amplifies normal estimation errors. Lack of benchmarks is also a challenge in this field. Almost all of the literature is evaluated on KITTI, which can be explained by the lack of datasets covering road grades beyond gentle slopes of a few degrees. This limits evaluation of robustness on steeper incline and decline scenarios.

We propose a ground normal vector prediction pipeline with late-stage fusion between a monocular camera and an IMU. A lightweight transformer-based matcher first provides robust point correspondences, from which an inter-frame homography is computed and decomposed to recover the ground normal vector. An example output can be seen on Fig. 1. This normal is then constrained and smoothed via temporal filtering and finally rotated into a global frame using the IMU-derived camera pose, ensuring the estimate aligns coherently with real-world geometry. The method is then tested and evaluated on long stretches of sequences with varying road inclinations. Compared to our earlier publication [15], this work introduces three key improvements. Namely, a transformer-based matcher that enhances feature correspondence on low-textured roads, a SLERP-based [16] filtering that improves temporal consistency and mitigates drift in ground normal prediction, and a new dataset that captures a wide range of road inclines and declines for comprehensive evaluation. Building on these improvements, our main contributions are:

- We propose a robust and publicly available method for ground normal prediction, combining advanced transformer-based feature matching and SLERP-based temporal filtering for improved reliability.
- We standardize the evaluation protocol and release a novel, comprehensive dataset consisting of synchronized camera, LiDAR and IMU-derived pose data, specifically captured to enable rigorous benchmarking of ground normal prediction methods.
- We conduct extensive validation of the proposed method across diverse incline and decline scenarios in a continuous setting, which to our knowledge has not been explored before, and demonstrate improved accuracy and temporal stability compared to existing approaches.

II. RELATED WORK

Accurately estimating the ground plane normal of the road in front of a moving vehicle is a deceptively small sub-problem that underpins a wide spectrum of downstream tasks, including stable inverse perspective mapping [5], [6], [7], camera pose estimation [10] or 3D object detection [11]. Yet, despite its practical reach, the literature remains thin and fragmented. While several plane estimation methods exist, dedicated road normal estimation presents unique challenges, particularly on sloped and uneven terrain. A review of the key literature reveals significant diversity in methodology and, more importantly, a lack of standardized evaluation protocols and little consensus on what constitutes ground truth.

Our work is based on monocular camera prediction, favored for its low cost and widespread deployment in ADAS systems, there are a few methods based on other modalities worth mentioning. Various depth sensor-based (e.g. LiDAR) systems [17], [18], [19], [20], [21] and stereo camera algorithms [22], [23], [24], [25] produce dense, highly accurate 3D reconstructions. These high-fidelity outputs could provide an indispensable reference standard when developing and validating monocular camera-based algorithms. Our earlier work [15] and our current work utilize the LiDAR sensor to calculate a reliable and reproducible ground truth.

While monocular methods may be a little more sensitive to 3D estimation errors (e.g. ground normal, pose), their minimal hardware requirements and ease of integration make them highly attractive. Existing approaches to ground normal prediction can be categorized into geometry-based and learning-based methods, each possessing distinct advantages and limitations.

Earlier geometric methods leverage the assumption that the road can be locally approximated as a dominant planar surface, enabling efficient estimation of its orientation using homography-based techniques [26], [27], [28], [29]. These approaches offer simple formulations based on projective geometry, but often rely on sparse, manually selected features and lack mechanisms to ensure robustness in visually challenging environments. Notably, Dragon et al. [13] embedded the ground normal and the camera translation in a Hidden Markov Model (HMM) that enforced orthogonality

and temporal smoothness while sampling homographies from minimal point correspondences.

Learning-based methods aim to predict the ground plane directly from raw image data using neural networks [14], [30], [31], often trained in a supervised or self-supervised manner. These methods bypass explicit geometric modeling in favor of data-driven representations, achieving strong performance when adequate training data is available. GroundNet [14] developed a supervised single-frame estimation method by jointly predicting depth, surface normals and ground segmentation, then enforcing a consistency loss between the two normals obtained via depth-to-plane fitting and direct normal regression. Its ground truth was simply the camera-to-vehicle extrinsic, i.e. the normal beneath the sensor, which could fail on sharp turns or rapidly changing slopes. Subsequent works moved toward self-supervision. Road Aware [30] learns depth, homography and ground normal jointly by minimizing photometric consistency between adjacent frames under a flat-road constraint, requiring only semantic masks for LiDAR plane extraction at training time. Xiong et al. [31] go one step further and jointly learn depth, pose, ground segmentation and normal without any labels, coupling the tasks through three geometric losses. Although promising, both methods inherit the dataset bias of KITTI's mostly flat urban drives.

Building on these developments, Zhang et al. proposed an odometry-based approach that relies on ego-motion measurements rather than dense image features. They use an Invariant Extended Kalman Filter (IEKF) to model the coupling between odometry and the ground plane, and the residual rotation between predicted and measured poses directly yield the ground normal vector [32]. Although the IEKF is decoupled from the odometry source, its accuracy remains dependent on the fidelity of the input motion estimates.

While both classical and learning-based homography pipelines have seen significant progress, three key challenges remain unaddressed at scale. Robustness in low-textured environments often fails when relying on sparse or handcrafted feature detectors. Temporal drift persists due to simplistic filtering methods, and no standardized, large-scale benchmark exists that covers realistic variations in road slope and surface conditions.

In our work, we revisit classic formulations with a modern pipeline that incorporates robust feature matching and temporal filtering, allowing accurate road normal estimation even in the presence of noise, slope variation, and texture-poor scenes. Moreover, our work is extended with a rigorously defined evaluation protocol applied to our newly collected, slope-diverse dataset. The following section describes how each component of our pipeline builds upon and significantly extends the preliminary results presented in [15].

III. PROPOSED METHOD

Our goal is to recover the three-dimensional orientation of the road surface, expressed as a unit normal vector, using only a monocular camera and an IMU-derived pose estimate. The core of our method is that at each timestep, we assume that we have access to a reliable camera pose (position and orientation) in a global frame, which is provided by the mentioned pose estimation. This pose is then combined with the road normal prediction via late-stage sensor fusion to accurately predict the road inclination ahead. Fig. 2 presents the complete pipeline. The subsequent description follows the stages depicted in this figure. To aid the reader, a numerical example tracing the full pipeline from a raw homography matrix through intrinsic normalization, decomposition into $\mathbf{R}$, $\mathbf{t}$, and $\mathbf{n}$, and the final normal vector extraction is provided in the supplementary material.

The algorithm's main part is the monocular camera pipeline, which takes a pair of successive frames as input from an image stream, current frame i_2 and previous frame i_1. During preprocessing, Adaptive Histogram Equalization (AHE) [33] is applied to enhance contrast and emphasize subtle road textures.

The processed frame pair is received by our homography computation module. Here we have chosen to integrate a lightweight transformer-based matcher called Efficient-LoFTR [34], a high-performance variant of the original LoFTR model [35]. This matcher algorithm can directly predict dense correspondences without requiring explicit keypoint detection or descriptor matching, unlike ORB or SIFT-based matchers. It also maintains a strong performance on low-textured surfaces which significantly enhances feature matching robustness while also being computationally efficient. The resulting correspondence map is subsequently passed to our homography estimation module. Here we adopt MAGSAC [36], a relatively new robust model estimation technique that improves upon RANSAC by marginalizing over a range of noise scales instead of relying on a fixed inlier threshold. While RANSAC classifies correspondences as inliers or outliers based on a hard threshold, MAGSAC evaluates model quality by integrating over possible noise scales, resulting in smoother and more statistically grounded estimations.

The homography calculations are constrained to the road ahead to reduce the influence of non-road parts of the image. The pre-defined region-of-interest (ROI) is defined based on a distance to make timely predictions possible by extending it beyond the expected reaction distance as determined by the assumed speed. This is currently at the 6 m mark.

After successful calculation, the resulting stable homography is passed on to the normal vector extraction module, where it is normalized to remove the effects of the camera's intrinsic parameters, improve numerical stability and enhance the interpretability of the transformation by isolating the geometric information in the scene. This is accomplished by pre- and post-multiplying the homography matrix with the inverse and the original intrinsic matrix respectively. If the unnormalized homography is denoted as

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}, \tag{1}$$

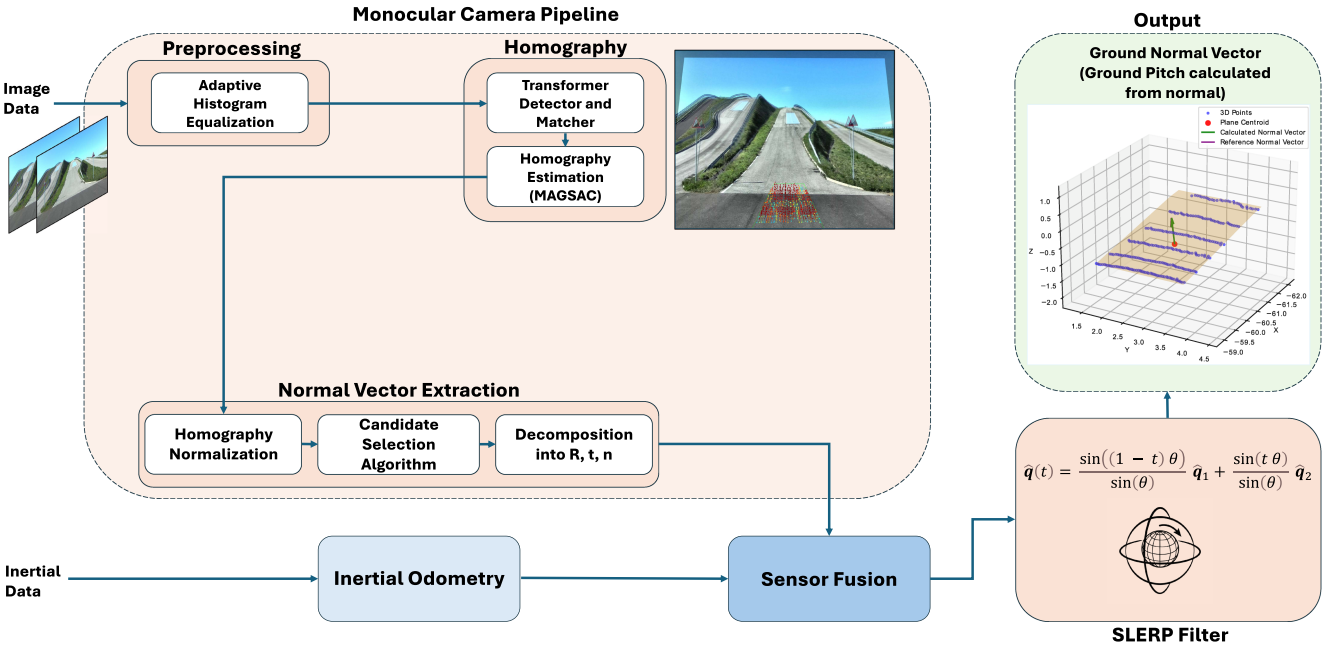

FIGURE 2. Overview of our algorithm. The monocular camera pipeline takes the previous and the present image, processes it, then fuses it with the inertial data. The final output is returned after the SLERP filter interpolation.

then the intrinsic normalization is performed by computing

$$\mathbf{H}_I = \mathbf{K}^{-1}\,\mathbf{H}\,\mathbf{K}, \tag{2}$$

where K represents the camera's intrinsic matrix

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

This matrix describes the mapping from metric camera coordinates to pixel coordinates, encoding the focal lengths f_x, f_y in pixel units and the principal point (c_x, c_y).

Following this, the matrix is scaled so that its bottom-right element equals one. Denoting the element in the i-th row and j-th column of $\mathbf{H}_I$ as h_{ij}, we define $c = h_{33}$, the element in the third row and third column. Dividing by c ensures that the resulting matrix has a unit value in its bottom-right position, which is a prerequisite for the subsequent decomposition into rotation, translation, and plane-normal components. If we denote $c = H_{I,33}$ as discussed, then the final normalized homography matrix is given by

$$\mathbf{H}_N = \frac{1}{c}\mathbf{H}_I = \begin{bmatrix} \frac{h_{11}}{c} & \frac{h_{12}}{c} & \frac{h_{13}}{c} \\ \frac{h_{21}}{c} & \frac{h_{22}}{c} & \frac{h_{23}}{c} \\ \frac{h_{31}}{c} & \frac{h_{32}}{c} & 1 \end{bmatrix}. \tag{4}$$

The normalized homography can be decomposed into its rotation, translation and plane-normal components and is given by

$$\mathbf{H} = \mathbf{R} + \frac{\mathbf{t}\,\mathbf{n}^\top}{d},$$

where $\mathbf{R}$ is the 3×3 rotation matrix between the two camera poses, $\mathbf{t}$ is the 3×1 translation of the camera center expressed in the camera's coordinate frame, $\mathbf{n}$ is the unit normal in that same frame and d is the signed distance from the camera centre to the plane along $\mathbf{n}$.

Because we constrain our feature matches to the road surface, the recovered normal $\mathbf{n}$ directly represents the ground-plane orientation in the camera frame. Note that this decomposition only yields the ratio $\mathbf{t}/d$ rather than $\mathbf{t}$ and d individually and that enforcing positive depth among other constraints resolves mathematical ambiguity. To resolve these mentioned ambiguities, we have defined a number of constraints to eliminate inconsistent solutions with the physical world while at the same time allowing for various incline and decline grades. The constraints and algorithm outputs adhere to standard computer vision conventions, visualized on Fig. 3. A right-handed coordinate system in which the positive Z-axis points forward along the camera's optical axis, the positive X-axis points to the right, and the positive Y-axis points in the downward direction.

Late-stage fusion with odometry processed IMU data is crucial for anchoring our derived normal vector in a physically consistent frame and to improve our algorithm's robustness. While our homography criteria reliably extract a road plane normal in the camera frame, those relative measurements alone cannot resolve drift, scale ambiguity, or slow bias changes over time. By deferring IMU fusion until after homography estimation, we leverage the complementary strengths of both sensors. Transforming the results into the world frame contextualizes the results, resulting in cues that might not be apparent with relative measurements alone, thus

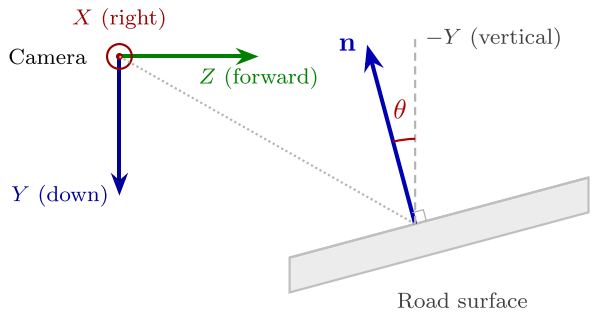

FIGURE 3. Camera coordinate convention and road surface normal geometry. The right-handed frame has its Z-axis along the optical axis (forward), X-axis pointing right, and Y-axis pointing downward. The unit normal vector $\mathbf{n}$ is perpendicular to the road surface and directed away from it. θ is computed from $\mathbf{n}$ via (7), corresponds to the angular deviation of $\mathbf{n}$ from the vertical direction ($-Y$); on a flat road $\theta = 0$ and $\mathbf{n}$ aligns with $-Y$.

enhancing the filtering process and improve our outlier rejection process. Additionally, it improves interpretability of the results, facilitating easier validation.

To improve on our previous linear Kalman Filter and to ensure temporal consistency, we integrated a Spherical Linear Interpolation (SLERP) filter into our pipeline [16]. The detailed operation of the fusion and temporal filtering stages is illustrated in Fig. 4. This algorithm provides numerically stable smoothing of the ground normal with low computational complexity, making it suitable for real-time filtering even on embedded platforms. In addition, it avoids the parameter sensitivities commonly encountered in Kalman Filter-based approaches.

At each new frame, the previous ground normal vector is expressed as a unit quaternion $\hat{\mathbf{q}}_1$, which is smoothly rotated toward the newly observed normal prediction, also represented as a quaternion $\hat{\mathbf{q}}_2$, using a fixed interpolation fraction $t \in [0, 1]$. It essentially computes the shortest path interpolation on the 4D unit hypersphere between the two quaternions (in this case our normal vectors) as

$$\hat{\mathbf{q}}(t) = \frac{\sin\left((1-t)\,\theta\right)}{\sin(\theta)}\,\hat{\mathbf{q}}_1 \;+\; \frac{\sin\left(t\,\theta\right)}{\sin(\theta)}\,\hat{\mathbf{q}}_2, \tag{5}$$

where the value of theta is

$$\theta = \cos^{-1}\left(\hat{\mathbf{q}}_1 \cdot \hat{\mathbf{q}}_2\right), \tag{6}$$

and the unit-quaternions are normalized as

$$\hat{\mathbf{q}}_i = \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|}, \quad i \in \{1, 2\}. \tag{7}$$

The interpolation is illustrated geometrically in Fig. 5. SLERP is well suited for ground normal estimation tasks. Since interpolation occurs directly on the unit quaternion sphere, the resulting vectors remain normalized by design, eliminating drift and ensuring physically valid outputs. Tuning the gain-like smoothing parameter t offers an intuitive trade-off between responsiveness and smoothness, controlling how far the estimate moves toward the new observation along the spherical arc. In contrast, Kalman Filter-based algorithms require careful tuning of process and measurement noise covariance matrices to perform well.

The filter's output provides the final refined ground surface normal, from which we derive the corresponding pitch angle

$$\theta = \arctan 2\left(-n_y, \sqrt{n_x^2 + n_z^2}\right) \tag{8}$$

over the same region to enable smooth, continuous plotting and better interpretability over time.

A. GROUND TRUTH

To anchor our performance claims and enable reproducible comparisons, a clear, well-defined ground truth is essential. The goal of the ground truth is to determine the vehicle's pose on the road ahead within the same region of interest used in our algorithm. This way, the ground truth is essentially limited only by the LiDAR sensor's measurement accuracy. We also publish the exact ROI boundaries and RANSAC [37] parameters, so any subsequent method can reconstruct the exact ground truth used in this paper.

Since our algorithm predicts from a specific monocular image region, we derive the ground truth by extracting the same region from the point cloud. This is done by transforming the LiDAR point cloud into the camera coordinate frame using the extrinsic matrix between the camera and the LiDAR sensor, then projecting the transformed points onto the image utilizing the intrinsic camera matrix. During this operation, each point's original index is preserved for the upcoming calculation. After projecting the LiDAR points into the image plane, the points whose projections lie within the ROI are selected. The original point cloud is filtered based on these projected points using their corresponding original indices. The RANSAC plane fitting algorithm is then applied to this filtered cloud. It samples minimal sets, fits planes and retains the model with the most inliers. The best fit plane is used to calculate the ground truth normal vector for the corresponding image segment.

The exact algorithm uses RANSAC plane fitting with a distance threshold of 0.01 m and a maximum of 1000 iterations. Before RANSAC, we remove outliers via Local Outlier Factor with 50 neighbors, 1% contamination, using the Euclidean distance metric. Finally, a temporal spike-suppression filter maintains a running window of 5 normals and rejects any new normal whose x-component (the component perpendicular to the plane) deviates by more than 0.06. A runtime example of the resulting reference and predicted normal vectors on the fitted plane is shown in Fig. 6.

The RANSAC distance threshold of 0.01 m reflects the sub-centimeter range accuracy of the Ouster OS1 LiDAR, ensuring the consensus set tightly approximates the true road surface. With 1000 iterations, the probability of sampling at least one outlier-free minimal set exceeds 99% even at moderate outlier ratios. The LOF parameters (50 neighbors, 1% contamination) follow established defaults for density-based outlier detection and proved effective at removing above-road objects and multi-path reflections without discarding valid

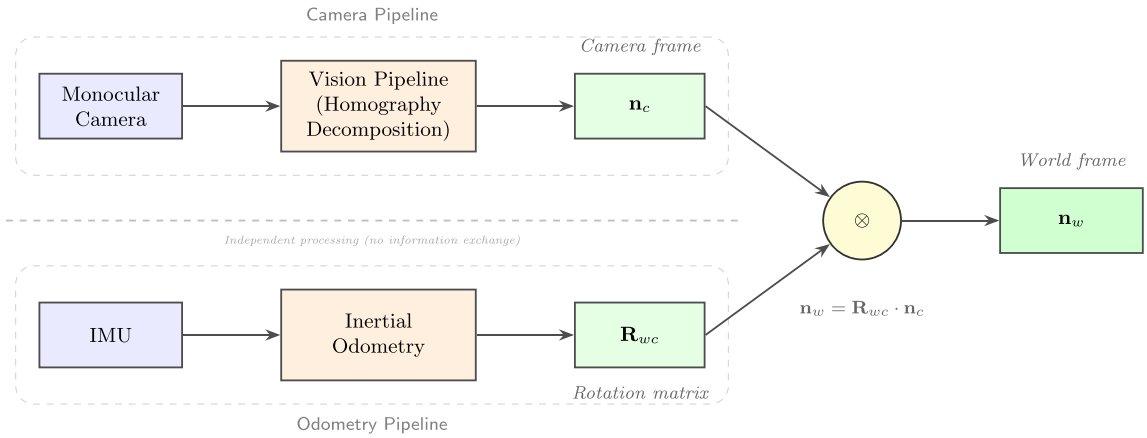

FIGURE 4. Late-stage sensor fusion architecture for road surface normal estimation. The camera and IMU pipelines operate independently, with fusion occurring only at the output stage through the coordinate transformation $\mathbf{n}_w = \mathbf{R}_{wc} \cdot \mathbf{n}_c$. This architecture maintains modularity and prevents error propagation between pipelines while enabling transformation of camera-frame normals to physically meaningful world-frame estimates.

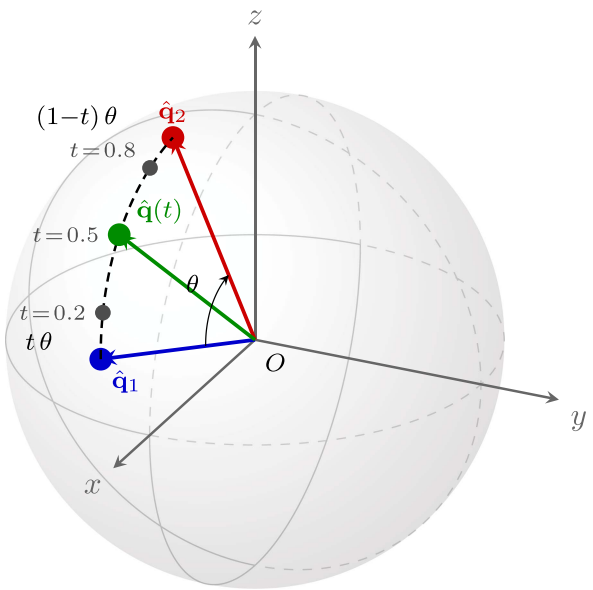

FIGURE 5. Geometric interpretation of SLERP on the unit sphere. The previous filtered estimate $\hat{\mathbf{q}}_1$ and the new observation $\hat{\mathbf{q}}_2$ are unit quaternions representing ground normal orientations, with $\theta = \cos^{-1}(\hat{\mathbf{q}}_1 \cdot \hat{\mathbf{q}}_2)$ denoting their angular separation from (5). The interpolated result $\hat{\mathbf{q}}(t)$ lies on the shortest great-circle arc defined by (4), with arc segments $t\,\theta$ and $(1-t)\,\theta$ corresponding to the weighting terms. Sample points at $t = 0.2$, 0.5, and 0.8 illustrate the smoothing trade-off: small t keeps the output near $\hat{\mathbf{q}}_1$ (stronger smoothing), while large t moves it toward $\hat{\mathbf{q}}_2$ (higher responsiveness). All interpolated results remain on the sphere surface by construction, preserving unit-norm validity without explicit renormalization.

road points. The spike-suppression window size of 5 frames and the 0.06 deviation threshold were determined empirically. More specifically, a shorter window provided insufficient context for detecting anomalies, while a longer one introduced excessive latency, and the 0.06 threshold corresponds approximately to a 3.4° change in surface orientation, which exceeds plausible frame-to-frame road geometry variation at typical driving speeds and frame rates.

Fitting a plane to dense LiDAR returns yields a stable, accurate road-normal estimate. However, with sparse point clouds common in mid-range LiDAR, the plane fit resolution can degrade. As [38] shows, temporal upsampling of consecutive LiDAR sweeps increases point density in the ROI, enabling a more precise planar fit. Combining multiple time aligned scans thus maintains LiDAR's resistance to transient vehicle noise while reducing sparsity, producing a more robust ground truth normal.

Our method is also compared to other papers, where possible. The exact method for these comparisons is detailed in Section IV.

IV. EXPERIMENTS

Ground normal estimation methods published so far lacked standardized evaluation protocols, particularly for long-range, dynamic driving scenarios involving significant road grade variation. To address this, we adopt a continuous evaluation strategy based on the three datasets we are going to introduce in Section A. This continuous evaluation allows both methods that are designed for continuous operation and those that process individual frames independently, regardless of temporal context, to be tested under the same protocol. The motivation and rationale behind this choice are discussed in detail in Section IV. To compare our results, we benchmark them against the approach proposed by Zhang et al. [32], which constitutes the previous state of the art for ground normal estimation. Their algorithm operates on continuous sequences as well, therefore offering a direct comparison. Our method is also compared to other papers, where possible. The exact method for these comparisons is detailed in Section IV as well.

The datasets we use (Section A) differ significantly in their characteristics, such as the magnitude and variability of road grades, the richness of surface textures, the length of continuous sequences and the resolution of the processed images. Evaluating algorithms across all three therefore demonstrates

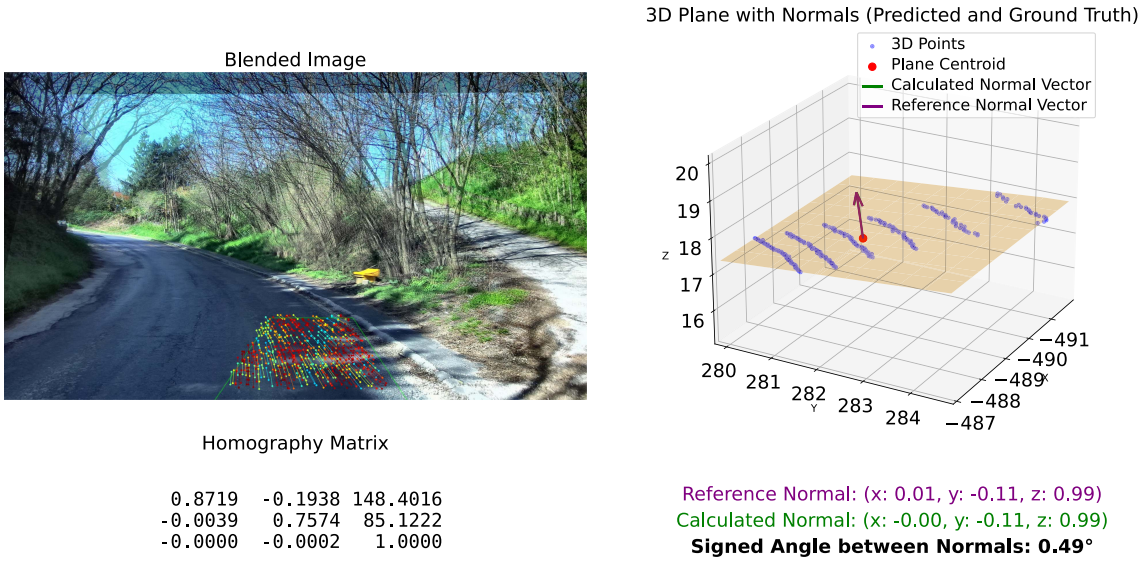

FIGURE 6. Snapshot of our method's output during runtime. The current frame demonstrates stable homography on a road lacking strong visual features. The image is from our own data collection (GradeSet).

their robustness to a wide range of real-world driving conditions. Together, the continuous testing protocol and the selection of complementary datasets ensure rigorous evaluation of our method's robustness across varied incline and decline scenarios and enable fair comparison of both geometry- and learning-based approaches over extended sequences. We report results using several metrics well established in the literature, that will be introduced in Section B.

As the basis of evaluations, we are using normal error, as it was done in previous literature already and calculating the pitch from the normal vector, using (8), which obtains the road pitch angle directly from the ground plane normal rather than extracting it as an Euler angle from the full orientation matrix. The normal-based formulation avoids the gimbal-lock singularity inherent to Euler parametrizations, requires only a single two-argument inverse tangent ($arctan2$) evaluation, therefore numerically better conditioned and it is a good approximation of the future pitch at the predicted region.

We adopt a LiDAR-based ground truth, computed on the subset of points that project into the same image-region of interest used by our monocular method (see Section A). By transforming each scan into the camera frame, preserving point indices, and selecting only those whose projections lie within the ROI, we ensure that the reference normal is computed from exactly the same spatial region that the algorithm processes. Moreover, because the procedure requires only extrinsic and intrinsic calibration plus a RANSAC fit, it can be applied to all three of our utilized datasets (Section A).

A. DATASETS

1) GRADESET DATASET

To conduct a more thorough validation of our method and to allow for a more robust benchmarking, we have recorded several sequences for the problem of ground normal prediction. We named our dataset GradeSet, which contains two main types of sequences: controlled incline trajectories that include a smooth transition over a rounded hilltop followed by a descent, and dynamic driving scenarios in which uphill and downhill segments appear together within a single continuous sequence. The controlled incline trajectories were recorded at the Hill Track facility of the ZalaZONE Automotive Proving Grounds, which offers test slopes with 5%, 12%, 18%, 25%, and 30% grades. A dedicated sequence was recorded for each incline. Several dynamic driving scenarios were also recorded in the hilly areas surrounding the city of Győr. There is a total of 14 sequences with varying length, sequences 8 through 12 are the hill sections, with increasing grades respectively, the rest are dynamic scenarios.

All sequences were captured using the same sensor setup and identical sensor configurations to ensure consistency across the dataset. Table 1 summarizes the specifications of each sensor, and Fig. 7 illustrates their placement on the data collection vehicle. The monocular images were captured using a ZED2i camera operating in 2 K mode with a 4 mm polarized lens, while the LiDAR data was recorded with a 64-channel Ouster OS1 sensor. The dataset includes the extrinsic calibration matrix between the camera and LiDAR to support accurate sensor fusion. Inertial measurements were obtained from a MicroStrain 3DM-GX5 IMU, which were processed by the LIO-SAM [39] odometry pipeline to produce accurate localization and pose estimates. As a secondary source of localization and pose information, GPS data is also provided. Software synchronization aligns the asynchronous sensor streams by selecting measurements with the closest timestamps, ensuring a maximum temporal offset below 80 ms. The sequences are publicly available, and we plan to extend the dataset in future work to include additional road

TABLE 1. Sensor Suite Specifications for the GradeSet Dataset

Parameter	Value
ZED2i Stereo Camera (left image used)	
Resolution	2208 × 1242 px (2K mode)
Lens	4 mm, polarized
Horizontal FOV	72°
Vertical FOV	44°
Aperture	f/1.8
Frame rate	10 Hz
Ouster OS1-64 LiDAR (Rev 06)	
Channels	64
Horizontal FOV	360°
Vertical FOV	45° (±22.5°)
Range	120 m (80% refl.) / 55 m (10% refl.)
Points per second	655,360
Frame rate	10 Hz
MicroStrain 3DM-GX5 IMU	
Output rate	100 Hz
Odometry backend	LIO-SAM [39]
Synchronization	
Method	Software (closest timestamp)
Max temporal offset	<80 ms

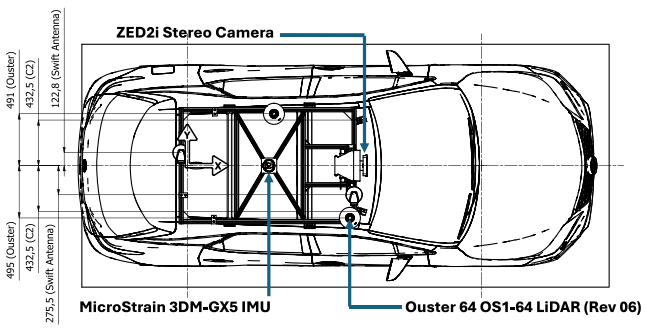

FIGURE 7. Sensor placement on the data collection vehicle used for GradeSet recordings. The forward-facing ZED2i camera and MicroStrain IMU constitute the online sensing suite used by the proposed pipeline. The Ouster OS1-64 LiDAR unit is mounted for ground truth generation only and is not used during inference. Measurements are in milimeters measured from the base link. The ZED2i and the MicroStrain are directly on the base link's x-axis. Sensor frames are described in this section in detail.

types, surface conditions, and driving scenarios. Moreover, there are already publications that discuss generating simulated data for the same tracks, extending the use cases of the dataset [40]. Each sensor operates in its own coordinate frame. The camera follows the standard computer vision convention, which is a right-handed frame with the Z-axis along the optical axis (forward), the X-axis pointing right, and the Y-axis pointing downward. The MicroStrain IMU and the Ouster OS1 body frame both follow a right-handed convention with the X-axis pointing forward, the Y-axis pointing left, and the Z-axis pointing upward. Note that the Ouster laser frame, in which raw point cloud data is natively expressed, uses an X-backward convention. The provided extrinsic calibration accounts for this distinction and transforms all measurements into a consistent reference frame prior to fusion.

2) PANDASET DATASET

PandaSet [41] was released in 2021 by Hesai and Scale AI and it contains over one hundred contiguous urban driving scenes recorded at 10 Hz from a Chrysler Pacifica outfitted with a 360° Hesai Pandar64 LiDAR, a forward-facing Hesai PandarGT LiDAR, six 1920 × 1080 resolution cameras and integrated GPS/IMU. The dataset comprises more than 48 000 images and 16 000 LiDAR sweeps annotated for cuboid detection and semantic segmentation across 28 classes. Each eight-second (80 frame long) sequence traverses various urban scenarios throughout the city of San Francisco, delivering pronounced incline and decline scenarios. PandaSet's rich sensor suite and annotations enable detailed error analysis on different surface types and challenging grades. It complements our research by enabling evaluation on steeper slopes and complex urban geometries. However, the presence of consistently rich road markings makes it less suitable for testing performance on low-textured road surfaces, and the relatively short duration of each sequence limits its utility for assessing long-term temporal stability.

3) KITTI DATASET

The KITTI Vision Benchmark Suite, introduced by Geiger et al. [42] in 2012, comprises 22 contiguous sequences captured at a resolution of 1241 × 376pixels. It is synchronized with 3D point clouds from a Velodyne HDL-64E LiDAR yielding roughly 100 000 points per frame and OXTS RT 3003 GPS/IMU data for precise localization and odometry ground truth. Each sequence averages 1.78km in length and the full suite covers approximately 39.2km of urban driving, offering varied road textures and gentle slopes. The dense stereo imagery and high-resolution LiDAR support baseline validation of monocular depth and normal estimation. KITTI's emphasis on relatively flat road segments makes it less ideal for evaluating ground normal estimation on steep grades. However, its long, contiguous trajectories provide usable test cases for assessing temporal consistency and performance under typical urban driving conditions. Compared to PandaSet and GradeSet, KITTI lacks significant road slope variation, but its established role as a widely used benchmark motivates its inclusion in our evaluation. A more detailed discussion of the limitations of the existing KITTI evaluation protocol is provided in Section IV.

4) KITTI EVALUATION PROTOCOLS

Evaluation of ground normal vector prediction on the KITTI benchmark presents unique challenges, which we aim to

clarify in this section for a clearer understanding. Currently, two mutually incompatible protocols have taken hold in the literature. The first, which we refer to as the odometry split, contains eleven odometry sequences (00-10) from the KITTI odometry benchmark that carry high-frequency 6-DoF pose measurements. Most works utilizing this split adopt either a 00-07 train, 08 validation, 09-10 test partitioning or simply test on 00-08 if the algorithm is not learning-based [13]. By leveraging vehicle ego-motion, methods can perform multi-frame fusion and enforce metric consistency, which is especially valuable when resolving subtle road inclines and declines.

In contrast, the Eigen et al. [43] split uses the entire KITTI dataset (KITTI raw), which does not contain 6-DoF pose information. It draws approximately 20000 unique training images from 28 raw sequences and evaluates on 697 sparsely sampled frames selected across the remaining 28 sequences, with no available pose information [44], [45]. It was originally introduced for monocular depth research and the exact sequences for the split are not published in Eigen et al. [43] with the process. However, after a repository release in 2016 for a publication by Garg et al. [46], a reconstructed list of training and test files was adopted and has since been used by the community. Some works, such as Wei et al. [2], introduced additional modifications to the test set, such as applying a planarity filter that removes frames with low RANSAC inlier ratios.

While the Eigen et al. split proved instrumental for early monocular depth research, it has its limitations for ground-normal prediction. First, the absence of pose measurements rules out any method from exploiting ego-motion cues, which could be critical for accurate ground-plane orientation prediction. Second, its 697 test images are non-contiguous snapshots sampled roughly every 5–10 meters.

This sparse, non-sequential sampling poses a conceptual limitation. Ground normal prediction benefits from temporal continuity because it allows subtle changes in road slope and orientation to be captured, outlier measurements to be filtered, and drift to be suppressed more effectively. For this reason, our method is designed from the ground up to process continuous image sequences. Namely, it estimates the homography from each pair of successive frames and the algorithm's temporal filtering also depends on an unbroken sequence. Even for methods that do not rely on consecutive frames, such sparse sampling dilutes the impact of systematic drift and local failure modes (for example sharp turns, inclines, dynamic lighting) that are only obvious in continuous operation. For a geometry task that must remain robust over a long stretch of time, this dilution could lead to over-optimistic error statistics. Also, 697 images represent less than 3 % of the frames available in the corresponding sequences. Our task, which explicitly targets inclines and declines, therefore needs a denser sampling to ensure adequate coverage of these scenarios.

The results are presented in Table 3. Our algorithm was tested on splits 00-08 and 09-10, since these dataset splits contain continuous frame sequences and pose information, which our pipeline requires. This choice enables us to test our algorithm in a fair way and also allows for sensor fusion between prediction and ego-motion provided in the odometry benchmark. Sequences 09 and 10 alone contain 2792 contiguous frames with synchronized 6-DoF pose, which is four times larger than the Eigen test split. The discussed protocols are summarized in Table 2.

TABLE 2. Summary of KITTI Evaluation Protocols (Dataset Splits) for Ground-Normal Prediction

Eigen *et al.* [43] (dataset) Split	
Test Sequences	28 sequences (KITTI raw), 697 frames
Pose Info	None
Frame Count	697
Sampling	Sparse (5-10 m intervals)
Ego-motion Use	Not allowed
Odometry (dataset) Split 00-08	
Test Sequences	00–08
Pose Info	6-DoF vehicle ego-motion
Frame Count	∼20.4 k
Sampling	Contiguous
Ego-motion Use	Allowed
Odometry (dataset) Split 09-10	
Test Sequences	09–10
Pose Info	6-DoF vehicle ego-motion
Frame Count	2792
Sampling	Contiguous
Ego-motion Use	Allowed

TABLE 3. Comparison of Ground Normal Prediction Algorithms on KITTI Evaluation Protocols

Eigen *et al.* [43] (dataset) Split	
Method	**Normal Error (deg)**
Xiong *et al.* [31]	3.02
Sui *et al.* [30]	1.12
GroundNet [14]	0.7
Zhang *et al.* [32]	3.26
Odometry (dataset) Split 00–08	
Method	**Normal Error (deg)**
Dragon *et al.* [13]	4.10
Zhang *et al.* [32]	4.53
Proposed Method (Ours)	**1.57**
Odometry (dataset) Split 09–10	
Method	**Normal Error (deg)**
Zhang *et al.* [32]	5.56
Proposed Method (Ours)	**0.97**

B. EVALUATION METRICS

We define a set of metrics that assess both the performance and interpretability of our algorithm and serve as the basis for the evaluation protocol. The two main metrics that we

use throughout the results section are the normal error and the pitch error, which is derived from the former. The normal error

$$E_{\text{normal}} = \frac{180}{\pi} \arccos\left(\bar{\mathbf{N}}_{\text{pred}} \cdot \hat{\mathbf{N}}_{\text{gt}}\right) \tag{9}$$

where the normalized vectors are given by

$$\bar{\mathbf{N}}_{\text{pred}} = \frac{\mathbf{N}_{\text{pred}}}{\|\mathbf{N}_{\text{pred}}\|}, \quad \hat{\mathbf{N}}_{\text{gt}} = \frac{\mathbf{N}_{\text{gt}}}{\|\mathbf{N}_{\text{gt}}\|} \tag{10}$$

is the most commonly used, which reports the mean angular deviation between the predicted and reference ground normal vectors. Here, $\bar{\mathbf{N}}_{\text{pred}}$ and $\hat{\mathbf{N}}_{\text{gt}}$ denote the estimated and ground truth 3D normals of the road surface, respectively, each normalized to unit length to compare orientation only. Using this as a baseline metric makes our algorithm directly comparable to prior work. Because a frame-by-frame visualization of the error trajectory is essential for rigorous benchmarking, we further compute the pitch error by converting each calculated and ground truth normal into a calculated (θ_i) and ground truth pitch angle ($\hat{\theta}_i$) using (8), and then taking the absolute difference

$$e_i = \left|\theta_i - \hat{\theta}_i\right| \tag{11}$$

Additionally this can also be considered the Mean Absolute Deviation Error, since averaging this quantity over all N frames yields

$$\text{MAE}_{\text{pitch}} = \frac{1}{N} \sum_{i=1}^{N} e_i \tag{12}$$

which is equivalent to the mean absolute error. The Root Mean Squared Error (RMSE)

$$\text{RMSE}_{\text{pitch}} = \sqrt{\frac{1}{N} \sum\nolimits_{i=1}^{N} e_i^2} \tag{13}$$

is also utilized to capture the impact of occasional large deviations. Because the squared differences give greater weight to outliers, RMSE is more sensitive than MAE to brief but pronounced errors and penalizes them more heavily.

In addition to the core error metrics, we also use a number of interpretability measures. They provide a complementary perspective on robustness that is essential for assessing the practical reliability and going beyond raw numerical scores. One of these metrics is the Angular Outlier Exceedance (AOE_3°), which quantifies outlier frequency as the percentage of frames whose per-frame pitch error e_i exceeds a threshold of 3° averaged over N frames. This metric is particularly useful to expose rare but critical outliers that are masked by aggregate statistics such as MAE or RMSE. Specifically, if you have N frames and you compute the per-frame error

$$e_i = \left|\theta_i - \hat{\theta}_i\right|, \tag{14}$$

then

$$\text{AOE}_3^\circ = \frac{|\{i \mid e_i > 3^\circ\}|}{N} \times 100\% \tag{15}$$

The complementary measure is Lag, which measures the temporal alignment between the estimated and ground truth pitch trajectories using cross-correlation, revealing how quickly the filter responds to changes in road slope.

$$r(\tau) = \frac{\sum_{i=1}^{N-|\tau|} \left(\theta_{i+\tau} - \bar{\theta}\right)\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)}{\sqrt{\sum_{i=1}^{N} \left(\theta_i - \bar{\theta}\right)^2} \sqrt{\sum_{i=1}^{N} \left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}} \tag{16}$$

where

$$\bar{\theta} = \frac{1}{N} \sum_{i=1}^{N} \theta_i \quad \text{and} \quad \bar{\hat{\theta}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i \tag{17}$$

are the means of the estimated and ground-truth pitch sequences respectively, and τ is the frame shift. The lag is then defined as the shift that maximizes this correlation

$$\text{Lag} = \arg\max_{\tau \in [-\tau_{\max},\, \tau_{\max}]} r(\tau). \tag{18}$$

A positive value indicates that the estimator reacts τ frames late. Multiplying by the frame period Δt converts the result to seconds if required. A concrete numerical example illustrating these metrics is provided in the supplementary material.

C. PERFORMANCE EVALUATION

Table 4 reports the comparative performance of our pipeline on PandaSet, KITTI (Seq. 09–10), and GradeSet against the previous state-of-the-art [32]. Based on these results, our method achieves lower errors and dramatically reduced lag, demonstrating clear, consistent improvements over the baseline. This uniform improvement highlights our algorithm's strong generalization capabilities. It performs almost equally well on PandaSet, which has dense road markings but relatively short sequences, on KITTI, with long stretches of predominantly flat terrain, and GradeSet, which features wide variations in road grade. Crucially, all of these results are obtained without any dataset specific recalibration. The same settings are used verbatim across PandaSet, KITTI, and GradeSet. The consistency of our improvements across such diverse driving domains underscores both the robustness and the reproducibility of our approach.

As described in Section B, we use pitch error to capture the qualitative behavior of our method. To assess performance across varied conditions, we present one pitch error plot for each dataset, selecting the sequence that best exemplifies challenging incline or decline scenarios.

We selected one of the most dynamic scenes from the KITTI dataset for evaluation. As illustrated in the pitch plot, the pitch angle on Fig. 8 (sequence 10) varies between approximately 10° and -5° over the course of more than 1200 frames. Although the overall road gradient is less steep compared to the other two examples, this sequence presents additional challenges due to frequent sharp turns and significant occlusions caused by narrow roadways, oncoming traffic, and parked vehicles. Despite these adverse conditions, our algorithm demonstrates strong performance on this sequence, achieving a mean normal error of 1.25° and a pitch error of 0.78°.

TABLE 4. Performance Comparison of Normal and Pitch Estimation Across Datasets. All Metrics are Computed Per-Frame and Averaged First Over Each Sequence, Then Across Sequences. The Evaluation on Several Datasets With Different Domains Showcases Our Method's Generalization Capabilities

Method	RMSE (°)	AOE_3 (%)	Lag (fr)	Normal Error (deg)	Pitch Error (deg)
PandaSet					
Zhang et al. [32]	0.83	2.91	4.45	1.68	0.62
Ours	**0.55**	**1.24**	**0.30**	**0.61**	**0.38**
KITTI (Sequence 09-10)					
Zhang et al. [32]	4.09	36.73	21.5	5.56	3.35
Ours	**0.88**	**1.46**	**3.0**	**0.97**	**0.61**
GradeSet					
Zhang et al. [32]	2.26	21.84	9.43	3.01	1.83
Ours	**1.09**	**1.36**	**2.29**	**1.18**	**0.74**

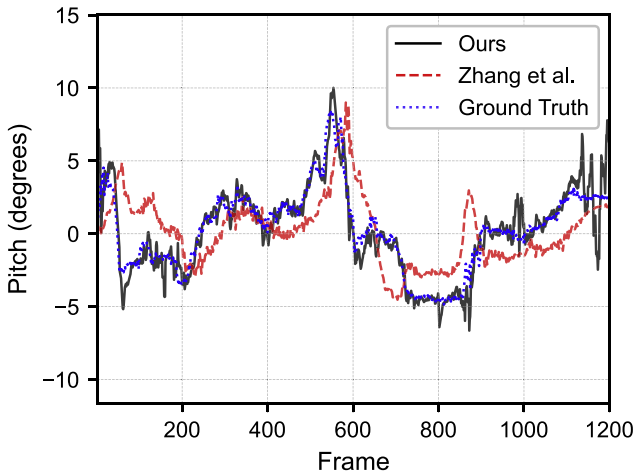

FIGURE 8. Comparison of pitch values on sequence 10 (KITTI dataset) against Zhang et al. [32]. We have chosen to highlight this sequence because it had the most grade variability among KITTI sequences.

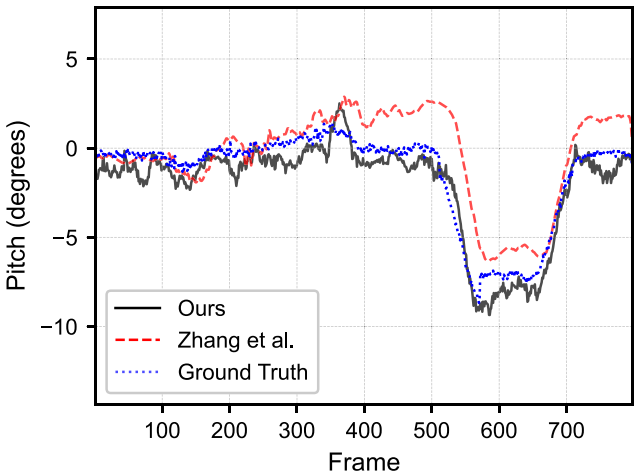

FIGURE 9. Comparison of pitch values on sequence 011 (GradeSet) against Zhang et al. [32]. The sequence is recorded on a 25% grade which was chosen to demonstrate the performance of the algorithm on a significant incline grade.

From our own dataset, we have chosen the 25% incline grade scenario (Fig. 9), which features a 25% steepness, effectively demonstrating the algorithm's performance on sudden inclines. The algorithm achieves a mean normal error of 1.25° and the pitch component is only 0.78° for the whole sequence, almost identical to the less steep KITTI sequence (Fig. 8), which demonstrates the method's robustness. It can also be observed that the reference method exhibits significant drift on this steep segment, whereas our approach remains relatively stable throughout. This further supports our claim that drift is a critical issue in steep hill scenarios and highlights the stability of our method under such conditions. It's important to note, that this scenario has virtually no road markings in front of the vehicle, demonstrating the transformer-based matcher's long-term stability.

PandaSet features significantly shorter sequences compared to the previous two datasets. As shown in Fig. 10, this example captures a more complex road incline within an urban environment. Despite the challenging conditions, our method demonstrates stable performance, with smooth transitions during the incline and smaller overshoots compared to the

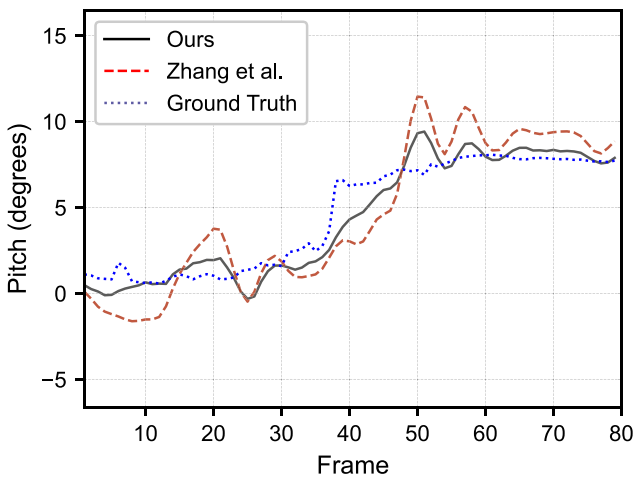

FIGURE 10. Comparison of pitch values on PandaSet sequence 029 with the method of Zhang et al. [32]. This sequence features a substantial change in road grade and includes a slight leftward turn, making it a representative example for evaluating performance under challenging urban conditions.

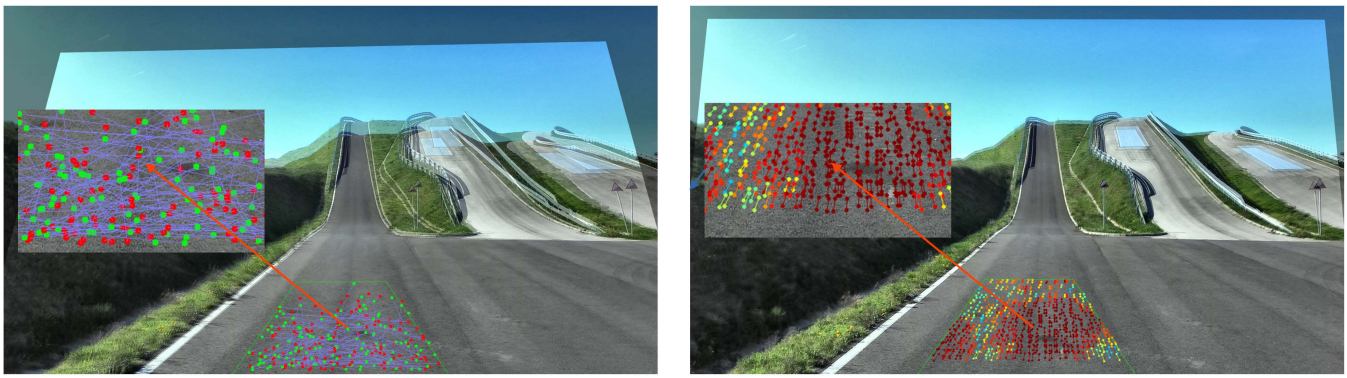

FIGURE 11. Comparison of the old SIFT-based matcher [47] (left), and the new EfficientLoFTR-based matcher [34] (right). Consecutive frames are blended on each other using the homography matrix, calculated from the respective visualized matches. It can be seen in the magnified regions, that the old matcher struggles to find correct matches, while the new, transformer-based matcher finds stable matches on a low-textured road surface. This is also evidenced by the smooth blend of the two consecutive image frames for the new matcher.

TABLE 5. Impact of Incremental Improvements Compared to Our baseline [15] Demonstrated on PandaSet

PandaSet			
Configuration	**MAE**	**RMSE**	**Normal Error (deg)**
Baseline [15]	0.57	0.82	1.15
Baseline [15] + Matcher	0.38	0.57	0.83
Proposed (Matcher + Filter)	**0.38**	**0.55**	**0.61**

reference method. The proposed method achieves a normal error of 1.18° and a pitch error of 0.75° on this sequence, compared to the reference method's errors of 3.15° and 1.72°, respectively.

Overall, it can be observed that despite all the challenging attributes, namely low-textured roads, steep inclines and declines, and sudden inclination changes, the algorithm outperforms all existing algorithms on continuous monocular sequences, thereby establishing a new state-of-the-art in the field of monocular ground normal estimation. Per-sequence results for all evaluated datasets are provided in Appendix A.

V. ABLATION STUDY

To quantify the individual and combined benefits of our proposed modules, we conduct an ablation study on PandaSet (Table 5). We adopt our previous algorithm from [15] as the baseline, both because our current work builds directly on it with multiple improvements and because it was originally evaluated on PandaSet, thereby ensuring a fair comparison. This study isolates each contribution, matcher and filter, showing how they improve or interact relative to our last algorithm. We begin with the original pipeline, which achieves an MAE of 0.57, RMSE of 0.82, and normal vector error of 1.15° on PandaSet.

With our feature matcher, we observe significant gains on roads with weak or repetitive textures. Fig 11 presents this improvement. One can see that the old SIFT-based matcher [47] struggles to find correct matches, while the transformer-based matcher on the bottom excels on the low-textured road. This

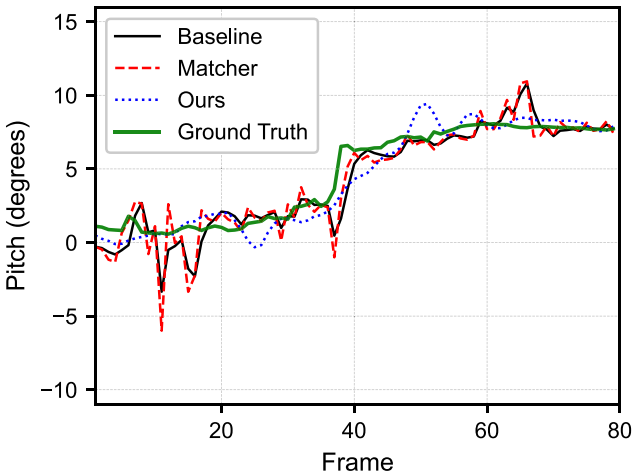

FIGURE 12. Ablation study components inspected on sequence 029 (PandaSet). Baseline denotes the preliminary pipeline [15] using SIFT-based matching and Kalman filtering. Matcher replaces only the feature matcher with EfficientLoFTR [34]. Ours denotes the full proposed pipeline combining EfficientLoFTR with the SLERP temporal filter. The sequence demonstrates the responsiveness of the matcher and the overall smoothness of the combined, final algorithm. Our exact errors for this sequence can be seen in Table 6. The mean ablation study results over the whole PandaSet are presented in Table 5.

can also be seen from the consistent flow of the matches forward, which was the direction of our measurement vehicle.

Table 5 shows that adding the matcher alone reduces MAE to 0.38, RMSE to 0.57, and normal error to 0.83°. These improvements stem from more reliable correspondence sets yielding better initial plane hypotheses in low texture regions.

If the temporal filter is added as well, we get the full, proposed pipeline, which leverages both matcher and filter. This version achieves the best results with a normal vector error of 0.61°. Fig. 12 plots this combined configuration against the LiDAR-based ground truth on the very same sequence illustrated in Fig. 10. In this plot, it can be seen that the matcher reacts almost promptly to genuine road-grade changes, although this swiftness comes with a trade-off, since

the algorithm is more susceptible to outliers this way. Together with both modules, the algorithm produces a smooth transient producing a stable estimate. This ablation confirms that the EfficientLoFTR matcher accelerates adaptation to new plane orientations, the SLERP filter enforces temporal consistency by rejecting outliers, and together they provide both agility and robustness far surpassing the baseline. Notably, the RMSE reduction from 0.57° to 0.55° when adding the filter, while the MAE remains at 0.38°, indicates that the filter's primary effect is suppressing occasional large deviations rather than shifting the central tendency of the error distribution.

VI. DISCUSSION

Based on the results and the ablation study, we share a few key insights from this work. The experimental results demonstrate that our AI and transformer-based feature matcher substantially enhances the homography-based ground plane estimation pipeline. While interframe homography estimation offers a conceptually simple solution, its real world performance has historically been limited by the lack of robustness of feature matching. In scenarios with sparse textures or repetitive road patterns, naive matching often fails, forcing downstream filters to handle wildly fluctuating predictions, resulting in filter overcompensation.

This lack of stability at the matching stage is precisely what our approach addresses. Our matcher delivers high quality correspondences even on sparse road features. This improvement simplifies the temporal filter's task, as evidenced by the ablation study (Table 5), where the combined configuration achieves the lowest normal vector and pitch angle errors without additional complexity.

A second source of stability comes from the way we smooth successive normal-vector estimates. SLERP is far simpler than a Kalman filter because it requires no covariance updates or matrix inversions while still enforcing unit-norm quaternions and constant rate rotation. These intrinsic constraints of the method align well with the gently changing orientation of typical road surfaces, even when the grade is steeper. This yields a more stable prediction stream with minimal computational overhead. In any case, the easy explainability of the algorithm ensures that the core geometric reasoning remains easily interpretable.

We would also like to highlight the cost effectiveness of our core contribution, which is the proposed monocular ground normal prediction pipeline. By developing a reliable camera-based pipeline for ground normal prediction, we can lower the entry cost for 3D terrain awareness without expensive sensors (e.g., LiDAR). This is particularly relevant for large-scale fleet deployment, where per-vehicle sensor cost is a decisive factor.

Despite the improvements demonstrated above, the proposed method has a few limitations that should be acknowledged. The homography-based pipeline assumes that the observed road surface within the region of interest can be approximated by a single plane. This assumption holds for continuous road segments but may produce unreliable estimates in scenarios where multiple distinct surfaces at different orientations contribute correspondences to the homography fit, such as curving hilltops or complex intersections. Additionally, although the transformer-based matcher significantly improves robustness on low-texture roads, completely featureless surfaces remain challenging. Vehicles, pedestrians, or other moving objects on the road surface may also contribute correspondences that violate the static-scene assumption underlying the geometric pipeline. While the robust estimation framework rejects outlier correspondences, scenarios with substantial dynamic object coverage in the region of interest may produce degraded estimates.

Several directions emerge for extending this work. The generalized applicability of the matcher opens up new possibilities for extending the normal vector estimation algorithm to several other fields. For instance, it could be adapted to non-structured and non-planar off-road scenarios, where feature distributions differ substantially. Moreover, these robust correspondences could enable reliable detection of curb edges or small gradients, improving applicability in urban settings. Beyond the core pipeline, installing multiple cameras could enable the algorithm to deliver side-view coverage with minor modifications. This would allow for accurate detection of ditches and curbs in close proximity, a capability that LiDAR-based sensor setups struggle to match without significant hardware additions. We are also planning on extending the entire pipeline with a coupled, AI-based ego-pose estimation module, enabling mutual refinement between pose and ground normal predictions through interconnected reasoning.

VII. CONCLUSION

This paper presented a monocular ground normal forecasting pipeline combined with late-stage IMU data fusion. It couples a transformer-based, robust feature matcher with SLERP-driven temporal filtering and outputs a reliable ground normal estimation, establishing a new state-of-the-art method. The late-stage fusion strategy delivers physically consistent and temporally stable ground plane normals on various incline and decline grades where other methods tend to drift. Beyond the raw accuracy gains, two contributions stand to benefit the wider research community: the publicly available algorithm code and the publicly released dataset to test ground normal estimation on extended sequences with sustained grade variations. Overall, the proposed method and the collected data move monocular terrain awareness and ground plane estimation closer to the reliability needed for large-scale, cost-effective deployment compared to LiDAR-based alternatives.

APPENDIX

A. ADDITIONAL EVALUATION RESULTS

We attached the per-sequence evaluation results for all three datasets used in this paper. All evaluated sequence results can be found in Table 7 for KITTI and in Table 8 for GradeSet. There are 47 test sequences overall for PandaSet, so we have chosen to include the five sequences that were chosen in our previous work [15]. PandaSet results are presented in Table 6.

TABLE 6. PandaSet Per-Sequence Normal and Pitch Errors (deg)

Seq.	Normal Error		Pitch Error	
	Zhang et al. [32]	Ours	Zhang et al. [32]	Ours
020	0.83	**0.38**	0.42	**0.25**
029	3.15	**1.18**	1.72	**0.75**
034	1.06	**0.84**	0.65	**0.43**
039	3.03	**0.88**	1.60	**0.69**
041	2.42	**0.89**	1.79	**0.76**
Average	1.68	**0.61**	0.63	**0.38**

TABLE 7. KITTI Per-Sequence Normal and Pitch Errors (deg)

Seq.	Normal Error		Pitch Error	
	Zhang et al. [32]	Ours	Zhang et al. [32]	Ours
00	7.25	**1.29**	4.07	**0.82**
01	4.19	**3.09**	2.64	**1.33**
02	6.98	**0.86**	5.28	**0.60**
03	3.39	**1.22**	0.96	**0.63**
04	1.11	**0.49**	0.51	**0.40**
05	5.08	**1.61**	2.41	**0.95**
06	6.07	**0.64**	5.75	**0.47**
07	4.09	**3.13**	2.97	**1.58**
08	2.65	**1.77**	1.56	**1.12**
09	6.52	**0.69**	4.61	**0.46**
10	4.60	**1.25**	2.09	**0.77**
Avg	4.72	**1.46**	2.98	**0.83**

TABLE 8. GradeSet Per-Sequence Normal and Pitch Errors (deg)

Seq.	Normal Error		Pitch Error	
	Zhang et al. [32]	Ours	Zhang et al. [32]	Ours
001	3.62	**1.38**	2.62	**0.75**
002	2.48	**1.19**	1.91	**0.83**
003	2.37	**1.10**	1.35	**0.63**
004	3.62	**1.28**	2.58	**0.64**
005	3.72	**1.23**	1.47	**0.83**
006	3.57	**1.12**	3.23	**0.75**
007	3.55	**1.12**	2.01	**0.74**
008	1.60	**0.85**	**0.41**	0.77
009	1.81	**1.14**	**0.65**	0.80
010	2.58	**1.16**	1.56	**0.67**
011	2.62	**1.25**	1.34	**0.78**
012	3.19	**1.39**	2.03	**0.76**
019	4.34	**1.38**	2.09	**0.86**
020	3.15	**0.97**	2.40	**0.58**
Average	3.02	**1.18**	1.83	**0.74**

REFERENCES

[1] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[2] P. Wei, G. Hua, W. Huang, F. Meng, and H. Liu, "Unsupervised monocular visual-inertial odometry network," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 325–332.

[3] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, 2007.

[4] J. Jeong and A. Kim, "Adaptive inverse perspective mapping for lane map generation with SLAM," in *Proc. 13th Int. Conf. Ubiquitous Robots Ambient Intell.*, 2016, pp. 38–41.

[5] L. Reiher, B. Lampe, and L. Eckstein, "A Sim2Real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–7.

[6] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 194–210.

[7] N. Marko, T. Sziranyi, and A. Ballagi, "Terrain depth estimation for improved inertial data prediction in autonomous navigation systems," in *Proc. 2023 IEEE Int. Automated Veh. Validation Conf.*, 2023, pp. 1–6.

[8] M. Irani and P. Anandan, "Parallax geometry of pairs of points for 3D scene analysis," in *Proc. 4th Eur. Conf. Comput. Vis.*, 1996, vol. 1064, pp. 17–30.

[9] D. Pfeiffer and U. Franke, "Efficient representation of traffic scenes by means of dynamic stixels," in *Proc. 2010 IEEE Intell. Veh. Symp.*, 2010, pp. 217–224.

[10] J. Košecká and W. Zhang, "Extraction, matching, and pose recovery based on dominant rectangular structures," *Comput. Vis. Image Understanding*, vol. 100, no. 2–3, pp. 274–293, 2005.

[11] X. Chen et al., "3D object proposals for accurate object class detection," in *Proc. 29th Conf. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[12] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 376–389.

[13] R. Dragon and L. Van Gool, "Ground plane estimation using a hidden Markov model," in *Proc. 2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4026–4033.

[14] Y. Man, X. Weng, X. Li, and K. Kitani, "GroundNet: Monocular ground plane normal estimation with geometric consistency," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2170–2178, doi: 10.1145/3343031.3351068.

[15] N. Marko, Z. Rozsa, A. Ballagi, and T. Sziranyi, "Robust road surface normal and pitch prediction via IMU-camera fusion," in *Proc. Adv. Concepts Intell. Vis. Syst.*, 2026, pp. 591–603.

[16] K. Shoemake, "Animating rotation with quaternion curves," in *Proc. 12th Annu. Conf. Comput. Graph. Interact. Techn.*, Jul. 1985, vol. 19, no. 3, pp. 245–254, doi: 10.1145/325165.325242.

[17] O. Gallo, R. Manduchi, and A. Rafii, "Robust curb and ramp detection for safe parking using the Canesta ToF camera," in *Proc. 2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2008, pp. 1–8.

[18] H. Yu, J. Zhu, Y. Wang, W. Jia, M. Sun, and Y. Tang, "Obstacle classification and 3D measurement in unstructured environments based on ToF cameras," *Sensors*, vol. 14, no. 6, pp. 10753–10782, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/6/10753

[19] S. Choi, J. Park, J. Byun, and W. Yu, "Robust ground plane detection from 3D point clouds," in *Proc. 14th Int. Conf. Control, Automat. Syst.*, 2014, pp. 1076–1081.

[20] W. Zhang, "LiDAR-based road and road-edge detection," in *Proc. 2010 IEEE Intell. Veh. Symp.*, 2010, pp. 845–848.

[21] M. W. McDaniel, T. Nishihata, C. A. Brooks, and K. Iagnemma, "Ground plane identification using LiDAR in forested environments," in *Proc. 2010 IEEE Int. Conf. Robot. Automat.*, 2010, pp. 3831–3836.

[22] Y. H. Lee, T. Leung, and G. G. Medioni, "Real-time staircase detection from a wearable stereo system," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 3770–3773. [Online]. Available: https://ieeexplore.ieee.org/document/6460985/

[23] T. Schwarze and M. Lauer, "Robust ground plane tracking in cluttered environments from egocentric stereo vision," in *Proc. 2015 IEEE Int. Conf. Robot. Automat.*, 2015, pp. 2442–2447.

[24] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2189–2199.

[25] S. Se and M. Brady, "Ground plane estimation, error analysis and applications," *Robot. Auton. Syst.*, vol. 39, no. 2, pp. 59–71, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889002001756

[26] J. Zhou and B. Li, "Homography-based ground detection for a mobile robot platform using a single camera," in *Proc. 2006 IEEE Int. Conf. Robot. Automat.*, 2006, pp. 4100–4105.

[27] J. Klappstein, F. Stein, and U. Franke, "Applying Kalman filtering to road homography estimation," in *Proc. ICRA 2007 Workshop, Planning, Perception Navigation Intell. Veh.*, 2007. [Online]. Available: https://hal.science/hal-04609872
[28] J. Arróspide, L. Salgado, M. Nieto, and R. Mohedano, "Homography-based ground plane detection using a single on-board camera," *IET Intell. Transport Syst.*, vol. 4, no. 2, pp. 149–160, 2010, doi: 10.1049/iet-its.2009.0073.
[29] M. Knorr, W. Niehsen, and C. Stiller, "Robust ground plane induced homography estimation for wide angle fisheye cameras," in *Proc. 2014 IEEE Intell. Veh. Symp.*, 2014, pp. 1288–1293.
[30] W. Sui, T. Chen, J. Zhang, J. Lu, and Q. Zhang, "Road-aware monocular structure from motion and homography estimation," 2021, *arXiv:2112.08635*.
[31] L. Xiong, Y. Wen, Y. Huang, J. Zhao, and W. Tian, "Joint unsupervised learning of depth, pose, ground normal vector and ground segmentation by a monocular camera sensor," *Sensors*, vol. 20, no. 13, 2020, Art. no. 3737. [Online]. Available: https://www.mdpi.com/1424-8220/20/13/3737
[32] J. Zhang, W. Sui, Q. Zhang, T. Chen, and C. Yang, "Towards accurate ground plane normal estimation from ego-motion," *Sensors*, vol. 22, no. 23, 2022, Art. no. 9375. [Online]. Available: https://www.mdpi.com/1424-8220/22/23/9375
[33] S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.
[34] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient LoFTR: Semi-dense local feature matching with sparse-like speed," in *Proc. 2024 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 21666–21675.
[35] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. 2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8922–8931.
[36] D. Barath, J. Matas, and J. Noskova, "MAGSAC: Marginalizing sample consensus," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10189–10197, doi: 10.1109/CVPR.2019.01044.
[37] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, doi: 10.1145/358669.358692.
[38] Z. Rozsa and T. Sziranyi, "Optical flow and expansion based deep temporal up-sampling of LiDAR point clouds," *Remote Sens.*, vol. 15, no. 10, 2023, Art. no. 2487. [Online]. Available: https://www.mdpi.com/2072-4292/15/10/2487
[39] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "LIO-SAM: Tightly-coupled LiDAR inertial odometry via smoothing and mapping," in *Proc. 2020 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5135–5142.
[40] Z. Szalay, "Critical scenario identification concept: The role of the scenario-in-the-loop approach in future automotive testing," *IEEE Access*, vol. 11, pp. 82464–82476, 2023.
[41] P. Xiao et al., "PandaSet: Advanced sensor suite dataset for autonomous driving," in *Proc. 2021 IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 3095–3101.
[42] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[43] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. 28th Conf. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.
[44] F. Aleotti, G. Zaccaroni, L. Bartolomei, M. Poggi, F. Tosi, and S. Mattoccia, "Real-time single image depth perception in the wild with handheld devices," *Sensors*, vol. 21, no. 1, 2020, Art. no. 15, doi: 10.3390/s21010015.
[45] X. Zhang, Y. Xue, S. Jia, and X. Pei, "CCDepth: A lightweight self-supervised depth estimation network with enhanced interpretability," in *Proc. 27th IEEE Int. Conf. Intell. Transp. Syst.*, 2024, pp. 64–69.
[46] R. Garg, V. K. B. G. G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 740–756.
[47] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.

NORBERT MARKÓ received the M.Sc. degree in electrical engineering with a specialization in automation systems from Széchenyi István University, Győr, Hungary, where he is currently working toward the Ph.D. degree. He is a Development Engineer with Machine Perception Research Laboratory (HUN-REN SZTAKI), Budapest, Hungary, and Research Engineer with Vehicle Industry Research Center, Széchenyi István University. He is actively involved in research projects related to autonomous vehicle perception. His interests include sensor fusion and deep learning, vision-based AI, autonomous vehicles, and vision-based applications of large language models.

ZOLTÁN RÓZSA (Member, IEEE) received the Ph.D. degree in vehicle and transportation engineering from the Budapest University of Technology and Economics, Budapest, Hungary, in 2020. He is currently with the Faculty of Transportation Engineering and Vehicle Engineering of Budapest University of Technology and Economics and Research Fellow with the Machine Perception Research Laboratory of the Institute for Computer Science and Control (SZTAKI), Budapest. His research interests include automated guided vehicles, machine vision, 3D recognition, and reconstruction.

ÁRON BALLAGI received the M.Sc. degree in mechanical engineering and the second M.Sc. degree in metallurgical engineering from the University of Miskolc, Miskolc, Hungary, in 1994 and 1996, respectively, the B.Sc. degree in economy management from Széchenyi István University, Győr, Hungary, in 1997, and the Ph.D. degree in informatics from Széchenyi István University, in 2015. He is currently an Associate Professor and Head of the Department of Automation and Mechatronics, Széchenyi István University. His research interests include computational intelligence, fuzzy control, fuzzy signatures, fuzzy communication, and autonomous robotics. He is an active member of several scientific societies, such as the John von Neumann Computer Society Robotics Section, Hungarian Fuzzy Association, and IEEE Robotics and Automation Society. He has numerous publications related to his research, including works on topics, such as self-driving cars, sign language recognition, and the application of autonomous aerial vehicles in transportation.

TAMÁS SZIRÁNYI (Life Senior Member, IEEE) received the Ph.D. and D.Sci. degrees from the Hungarian Academy of Sciences, Budapest, Hungary, in 1991 and 2001, respectively. In 2001, he was a Full Professor with Pannon University, Veszprem, Hungary, and Peter Pazmany Catholic University, Budapest, in 2004. Since 2006, he has been a Research Scientist with the Institute for Computer Science and Control (SZTAKI), Budapest, since 1992, where he leads the Machine Perception Research Laboratory. He is currently a Full Professor with the Budapest University of Technology and Economics, Budapest. He has more than 310 publications, such as 60 in major scientific journals and several international patents. His research interests include machine perception, pattern recognition, texture and motion segmentation, Markov Random Fields and stochastic optimization, remote sensing, surveillance, intelligent networked sensor systems, graph-based clustering, and digital film restoration. He has participated in several prestigious international (ESA, EDA, FP6, FP7, OTKA) projects with his research laboratory. He was the founder and past president (1997 to 2002) of the Hungarian Image Processing and Pattern Recognition Society. He was an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING (2003-2009), and he has been an Assistant and now Area Editor of *Digital Signal Processing* (Elsevier) since 2012. He was honored by the Master Professor Award in 2001 and ProScientia (Veszprem) Award in 2011 and Officers Cross by the President of Hungary in 2018. He is a fellow of the International Association Pattern Recognition (IAPR) and Hungarian Academy of Engineering both from 2008. He has been the member of Hungarian Academy of Sciences since 2022.