QuadEdge-VO: Stereo Visual Odometry from Quadruplet Edge Correspondences

Jue Han[1⋆], Chiang-Heng Chien[1], and Benjamin Kimia[1]

Brown University, Providence RI 02912, USA

Abstract. This paper introduces a fundamental paradigm shift in edge-based visual odometry (VO). Conventional edge-based methods typically rely on iterative, energy-based minimization of reprojection errors to align disparate edge maps—a process that is computationally intensive and highly sensitive to initial pose estimates. In contrast, we propose QuadEdge-VO, a novel stereo VO framework that leverages discrete geometric primitives termed "edge quads". By pairing stable stereo edge features across a temporal sequence, we demonstrate that a single edge quad reduces the six-degree-of-freedom (6-DoF) pose to a one-parameter family of solutions, while a pair of quads enables a direct, non-iterative solution for relative pose within a RANSAC loop. To address the combinatorial explosion inherent in edge matching, we introduce a secondary contribution: a cascaded filtering architecture. By enforcing a suite of geometric and photometric constraints, our system achieves 98% precision and recall for stereo correspondences and maintains high reliability across temporal sequences. This filtering reduces tens of thousands of subpixel-accurate edges to a sparse set of highly reliable quads (89% recall, 87% precision). Experimental results demonstrate that QuadEdge-VO qualitatively and quantitatively outperforms state-of-the-art traditional and learning-based methods on the ETH3D dataset, offering a robust and efficient solution for real-time motion estimation in challenging environments.

1 Introduction

Visual Odometry (VO) is the process of estimating camera motion by analyzing a sequence of images acquired over time. By identifying and tracking visual information across successive frames, VO recovers the relative transformation between observations, typically as incremental rotations and translations. It is a central component of many Simultaneous Localization and Mapping (SLAM) systems, where frame-to-frame motion estimates are used to maintain a map of the surrounding environment. However, VO is also frequently used as a standalone capability in applications where accurate motion estimation is required but explicit mapping is unnecessary.

VO has been applied in a wide range of systems with different camera modalities. Monocular VO infers motion from a single camera, estimating a trajectory

⋆ Paper approved as a Master's report for this author

only up to an unknown global scale factor. Stereo or RGB-D cameras, in contrast, provide depths through geometric triangulation or active sensing, allowing the scale of motion to be directly observed.

Traditional approaches to VO are broadly categorized into feature-based, direct, and learning-based methods. Feature-based methods estimate motion by extracting salient keypoints and establishing correspondences, typically followed by geometric pose estimation and nonlinear optimization, such as Levenberg–Marquardt [15]. While robust, their performance depends on distinctive textures; in weakly textured or repetitive environments, feature matching reliability degrades. To mitigate this, point features are sometimes augmented with lines or planes to provide additional geometric constraints. Direct methods, in contrast, minimize photometric or geometric error directly over image intensities [4]. By utilizing a large number of pixels, they operate in scenes where distinctive keypoints are scarce. However, their reliance on brightness constancy makes them sensitive to illumination and exposure variations. Furthermore, the underlying optimization problems remain highly non-convex and may converge to local minima if initialization is poor. More recently, learning-based methods [22] have emerged, using deep neural networks to infer motion directly from sequences. While they can handle challenging conditions like motion blur or dynamic illumination, they often require substantial training data, lack geometric consistency, and may generalize poorly outside their training distribution.

Edge-based Visual Odometry (EVO) estimates camera motion by detecting and tracking image edges, which correspond to strong spatial intensity gradients. These typically arise at boundaries where there are abrupt changes in illumination, surface orientation, or material properties. Because these discontinuities often coincide with meaningful geometric structure, edges provide informative cues about the spatial layout. Edges are particularly attractive because they are abundant in structured man-made environments where points are scarce, stable under illumination variations, and resilient to motion blur. Additionally, edge extraction is computationally inexpensive and highly parallelizable, and edges exhibit predictable covariances under geometric transformations. Finally, they remain detectable even in moderate scene dynamics, making them useful in environments with moving agents.

Despite these advantages, exploiting edges presents several fundamental challenges. Unlike isolated point features, edges extend along one-dimensional structures and are dense along their tangent direction, leading to a combinatorial explosion in possible matches. This elongated structure also introduces local and global ambiguity, as small displacements along the edge may produce nearly identical image evidence. Another challenge is depth ambiguity: an observed edge may correspond to different 3D structures at varying depths. For these reasons, many edge-based VO approaches have focused on RGB-D sensing, where measured depths help disambiguate correspondences and constrain the alignment problem [17, 28]. Such methods frequently employ distance-transform representations to facilitate robust 3D–2D edge alignment. Alternatively, a stereo camera can provide depth through triangulation to reduce correspondence ambiguity.

This paper presents a fundamental paradigm shift in edge-based VO. Conventional methods typically estimate relative pose through energy-based minimization of reprojection errors to align disparate edge maps. This iterative approach evolves an initial pose estimate toward optimal alignment, a framework motivated by the historical difficulty of establishing sufficient reliable edge correspondences to estimate pose directly. In contrast, we introduce a novel stereo VO method that constructs reliable "edge quads" by pairing stable stereo edge features across a temporal sequence, Figure 1(a). We demonstrate that a single edge quad reduces the six-degree-of-freedom (6-DoF) pose to a one-parameter family of solutions. A pair of quads is thus sufficient to solve for the relative pose within a RANSAC loop. By leveraging the geometric constraints inherent in this over-constrained quad pair, we prune the hypothesis space prior to validation, resulting in an estimation process that is both computationally efficient and robust.

A critical secondary contribution is a framework for managing the combinatorial explosion inherent in edge-pair matching. Robust EVO necessitates high-fidelity edge extraction; achieving subpixel accuracy in position and orientation typically yields tens of thousands of dense edges per frame, Figure 2(a). While sub-sampling is a common heuristic to maintain tractability, it often degrades pose estimation by discarding fine-grained geometric information. To circumvent this, we propose a cascaded filtering architecture that leverages geometric and photometric cues to tame the matching space. For stereo imagery, this approach enables near-unambiguous edge correspondence, achieving 98% precision and 98% recall. Even in temporal sequences which lack epipolar constraints, we maintain high recall and precision. Integrating these stereo and temporal correspondences into "quads" introduces further structural constraints that reduce the pool to a few thousand highly reliable candidates per stereo pair. With an 85% recall and 87% precision rate, this filtered set provides a robust foundation for successful RANSAC-based estimation.

2 Related Works

The integration of edges as primitives in VO pipelines offers a robust alternative to point-based systems, though establishing efficient correspondences remains a central challenge.

Distance-Field and Registration-Based Methods: A significant branch of edge-based VO focuses on 2D-3D registration. Kuse and Shen [9] utilized precomputed distance transforms (DT) in RGB-D imagery to optimize edge map distances for pose estimation (see also [21]). To address the limitations of DT in complex scenes, CannyVO [30] proposed Approximate and Oriented Nearest Neighbor Fields (ANNF/ONNF) to improve registration accuracy. More recently, EdgeVO [28] introduced a selection strategy for RGB-D cameras that prioritizes edges based on their specific contribution to motion estimation, effectively balancing accuracy with computational overhead. Other RGB-D approaches, such as RESLAM [17], EDVO [3], and ROEVO [11], further underscore the popularity of active depth sensing in disambiguating edge correspondences.

Hybrid and Monocular Edge Tracking: To mitigate the uncertainty of edge endpoints, several frameworks combine edge primitives with other modalities. Wang *et al.* [25] integrated edges into direct VO pipelines, while Edge SLAM [13] proposed a monocular pipeline using optical flow for edge point tracking, subsequently refining correspondences across three views. In the domain of visual-inertial odometry, PE-VINS [10] utilizes point features in conjunction with edges to improve robustness. Despite these advances, they often remain reliant on iterative refinement or auxiliary sensors to resolve geometric ambiguities.
Learning-Based and Stereo Matching: Recent developments have seen a surge in learning-based stereo matching, *e.g.*, Selected-Stereo [24] and MochaStereo [2]. These methods prioritize disparity estimation accuracy specifically on edges; however, they are generally constrained to rectified images and lack the flexibility for general motion. Other matching frameworks, such as [12], are limited by pixel-level accuracy, *e.g.*, Canny edges, falling short of the subpixel precision required for high-fidelity VO.

In the context of stereo VO, MAC-VO [16] utilizes a learned uncertainty model to select semi-sparse features for pose graph optimization. Other modern stereo systems [14] rely on photometry, or fuses geometric and photometric cues [6]. While stereo edge-based systems like [7, 26] have made progress, they typically inherit the iterative optimization overhead found in earlier registration-based work. In contrast, our QuadEdge-VO framework leverages discrete geometric quads to solve for pose directly, bypassing the need for active depth sensing or intensive energy minimization.

3 Problem Formulation

This paper aims at computing the relative pose of a moving stereo rig over time. The relative pose between the left and right cameras, $(\mathcal{R}_0, \mathcal{T}_0)$, is known and fixed over time. This means that a point Γ^- in the left camera coordinate and its corresponding point Γ^+ in the right camera are related by

$$\Gamma^+ = \mathcal{R}_0 \Gamma^- + \mathcal{T}_0. \tag{1}$$

It is assumed that the stereo rig acquires imagery at a certain frame rate, producing pairs of images at regular intervals of time, Figure 1(a). Let the relative pose of the left and right cameras at each frame pair (shown in red) with respect to the previous frame pair (shown in green) is $(\mathcal{R}^-, \mathcal{T}^-)$ and $(\mathcal{R}^+, \mathcal{T}^+)$. Since $(\mathcal{R}^+, \mathcal{T}^+)$ is derivable from $(\mathcal{R}^-, \mathcal{T}^-)$, the pose of the left camera coordinate system over time is the primary set of unknowns, denoted as $(\mathcal{R}, \mathcal{T})$ for brevity, with points only expressed in the left camera coordinate system denoted by Γ and $\overline{\Gamma}$, respectively, with

$$\overline{\Gamma} = \mathcal{R}\Gamma + \mathcal{T}. \tag{2}$$

The central concept underlying *edge-based stereo visual odometry* is the formation of a quadruplet of edge correspondences. An edge is expressed as a location and a unit tangent, *i.e.*, (γ^-, t^-), expressing an edge in the left camera at the current time. The corresponding stereo edge in the right camera is denoted by (γ^+, t^+). These edges at the next time frames are denoted by $(\overline{\gamma}, \overline{t}^-)$ and

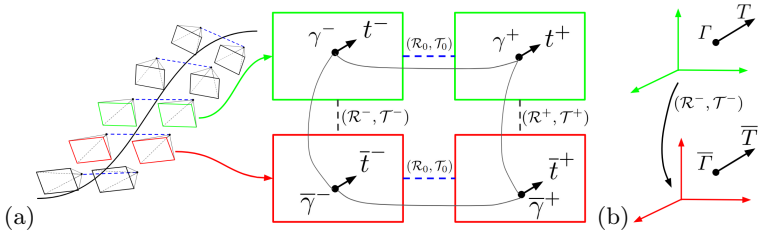


Figure 1: (a) Stereo edge odometry operates on sequential pairs of stereo images. The relative pose between cameras in the stereo rigs is known. The relative pose of a pair of the stereo rigs over time $(\mathcal{R}^-, \mathcal{T}^-)$ is not known. Our approach relies on formation of "quads", namely, two pairs of stereo edge correspondences. (b) The quad $(\gamma^-, \gamma^+, \overline{\gamma}^-, \overline{\gamma}^+)$ with associated unit tangents $(t^-, t^+, \overline{t}^-, \overline{t}^+)$ can equivalently be represented by the 3D reconstructions $(\Gamma, \overline{\Gamma})$ and the associated unit tangents $(T, \overline{T})$.

$(\overline{\gamma}^+, \overline{t}^+)$. This quadruplet of edges, $((\gamma^-, t^-), (\gamma^+, t^+), (\overline{\gamma}^-, \overline{t}^-), (\overline{\gamma}^+, \overline{t}^+))$, or in short "quad", is the basic building block of our algorithm, Figure 1(a).

Observe that in a stereo rig of known geometry $(\mathcal{R}_0, \mathcal{T}_0)$, the pair of corresponding edges $((\gamma^-, t^-), (\gamma^+, t^+))$ can be triangulated as Γ and the tangents can be reconstructed as well as T [29]. Similarly, the pair of edges $((\overline{\gamma}^-, \overline{t}^-), (\overline{\gamma}^+, \overline{t}^+))$ reconstruct as $(\overline{\Gamma}, \overline{T})$. Conversely, Γ and T project to each camera to give $((\gamma^-, t^-), (\gamma^+, t^+))$. Thus, a quad can be fully represented by $((\Gamma, T), (\overline{\Gamma}, \overline{T}))$, Figure 1(b). The next section explores whether a single quad is sufficient to find $(\mathcal{R}, \mathcal{T})$ and if not, how much additional information is required.

4 Using Quads for Moving Stereo Rig Pose Estimation

The first question to explore is to what extent a single quad, specified as $((\Gamma, T), (\overline{\Gamma}, \overline{T}))$, constrains the relative pose of a moving stereo rig.

Proposition 1 (Relative Pose from a Single Quad). *Let (Γ, T) and $(\overline{\Gamma}, \overline{T})$ be a pair of corresponding oriented points across cameras with unknown relative pose $(\mathcal{R}, \mathcal{T})$ related by*

$$\begin{cases} \overline{\Gamma} = \mathcal{R}\Gamma + \mathcal{T}, \\ \overline{T} = \mathcal{R}T. \end{cases} \tag{3}$$

Then, the set of admissible rotations $\mathcal{R}$ is a one-parameter family,

$$\mathcal{R}(\theta) = R_{\overline{T}}(\theta)\, R, \tag{4}$$

where $R_{\overline{T}}(\theta)$ denotes rotation by angle θ about axis $\overline{T}$,

$$R_{\overline{T}}(\theta) = \cos\theta\, I + (1 - \cos\theta)\, \overline{T}\, \overline{T}^\top + \sin\theta\, [\overline{T}]_\times, \tag{5}$$

and where R is the fixed rotation aligning T to $\overline{T}$ when $T \neq -\overline{T}$ given by

$$R = I + [v]_\times + \frac{[v]_\times^2}{1 + c}, \quad v = T \times \overline{T}, \quad c = T \cdot \overline{T}, \tag{6}$$

and $T = -\overline{T}$ *given by* $R = 2uu^\top - I$*, where* u *is any unit vector orthogonal to* T *and* $\mathcal{R}$ *defines the* $180°$ *rotation about axis* u*. For each such rotation* $\mathcal{R}(\theta)$*,*

$$\mathcal{T}(\theta) = \overline{\Gamma} - \mathcal{R}(\theta)\Gamma. \tag{7}$$

The proof of Proposition 1 is given in the supplementary material.

This proposition clearly states that with a single quad, the space of rigid motions is defined up to a single variable, giving a one-parameter family of solutions. Intuitively, the quad defines 5 of the 6 pose parameters through 3 equations from the $(\Gamma, \overline{\Gamma})$ correspondence and 2 equations from the $(T, \overline{T})$ correspondence. Additional information is needed to find the pose.

The smallest piece of information that can be added is to provide a secondary tangent correspondence. This could come at a *corner*, for example, where the point has two distinct orientations, T_1 and T_2, then a correspondence with $\overline{\Gamma}$ and $(\overline{T}_1, \overline{T}_2)$ would over-constrain the pose as the additional orientation correspondence provides two more equations. In the supplementary material, we show that this is a minimal problem with a constraint.

Corners are typically easy to find, but their orientations are difficult to determine. This motivates an alternative approach of using *two quads*, $((\Gamma_1, \overline{\Gamma}_1)$, $(T_1, \overline{T}_1))$ and $((\Gamma_2, \overline{\Gamma}_2)$, $(T_2, \overline{T}_2))$.

Proposition 2 (Relative pose from two oriented points). *Let two oriented points* (Γ_1, T_1) *and* (Γ_2, T_2) *be in correspondence with* $(\overline{\Gamma}_1, \overline{T}_1)$ *and* $(\overline{\Gamma}_2, \overline{T}_2)$*, respectively, such that*

$$\begin{cases} \overline{\Gamma}_1 = \mathcal{R}\Gamma_1 + \mathcal{T} \\ \overline{\Gamma}_2 = \mathcal{R}\Gamma_2 + \mathcal{T} \end{cases} \quad \textit{and} \quad \begin{cases} \overline{T}_1 = \mathcal{R}T_1 \\ \overline{T}_2 = \mathcal{R}T_2. \end{cases} \tag{8}$$

Then, a solution $(\mathcal{R}, \mathcal{T})$ *exists if and only if:*

$$\begin{cases} 1.\ \|\Gamma_2 - \Gamma_1\| = \|\overline{\Gamma}_2 - \overline{\Gamma}_1\| \\ 2.\ (\Gamma_2 - \Gamma_1) \cdot T_1 = (\overline{\Gamma}_2 - \overline{\Gamma}_1) \cdot \overline{T}_1 \\ 3.\ (\Gamma_2 - \Gamma_1) \cdot T_2 = (\overline{\Gamma}_2 - \overline{\Gamma}_1) \cdot \overline{T}_2 \end{cases} \quad \textit{and} \quad \begin{cases} 4.\ T_1 \cdot T_2 = \overline{T}_1 \cdot \overline{T}_2 \\ 5.\ T_1 \not\parallel (\Gamma_2 - \Gamma_1) \\ 6.\ \det[(\Gamma_2 - \Gamma_1), T_1, T_2] \\ \qquad = \det[(\overline{\Gamma}_2 - \overline{\Gamma}_1), \overline{T}_1, \overline{T}_2]. \end{cases} \tag{9}$$

If the vectors $((\Gamma_2 - \Gamma_1), T_1, T_2)$ *are not coplanar and* $T_1 \not\parallel (\Gamma_2 - \Gamma_1)$*, then the solution is unique and given by*

$$\mathcal{R} = \overline{B}B^\top, \qquad \mathcal{T} = \overline{\Gamma}_1 - \mathcal{R}\Gamma_1,$$

where B *and* $\overline{B}$ *are orthonormal frames constructed from* $(\Gamma_2 - \Gamma_1, T_1, T_2)$ *and* $(\overline{\Gamma}_2 - \overline{\Gamma}_1, \overline{T}_1, \overline{T}_2)$*.*

The proof of Proposition 2 is provided in the supplementary material.

5 Formation of Quads

The formation of all potential quads of edges $((\gamma^-, t^-), (\gamma^+, t^+), (\overline{\gamma}^-, \overline{t}^-), (\overline{\gamma}^+, \overline{t}^+))$ is combinatorially overwhelming: There are typically tens of thousands of edges per image: an image of size 942×489 has on average about $N = 50,000$ edges, Figure 2. The brute force approach of combining edges to form quadruplets leads to a combinational explosion and quickly becomes computationally infeasible. Fortunately, the pairing of (γ^-, t^-) and (γ^+, t^+), as well as the pairing of $(\overline{\gamma}^-, \overline{t}^-)$ and $(\overline{\gamma}^+, \overline{t}^+)$ are both limited to those satisfying the epipolar constraint. This cuts down the numbers of each pairing by a factor of two, assuming the area of the epipolar band is 1/100 of the area of the image. Nevertheless, the number of hypotheses remains vast and other cues are necessary to guide the formation of viable stereo edge pairs without dismissing a large percentage of the veridical pairings. This is discussed below for pairing edges across stereo images. Fortunately, all cues except the epipolar constraint are also applicable to pairing edges over time, namely the pairing of (γ^-, t^-) and $(\overline{\gamma}^-, \overline{t}^-)$, and the pairing of (γ^+, t^+) and $(\overline{\gamma}^+, \overline{t}^+)$. Finally, the circular constraint of formaing a quad helps reduce possibilities, as discussed below.

5.1 Pairing edges across stereo images

Each edge in one of the stereo image pair has either no correspondence in the other (say, due to occlusion), or its correspondence is unique. A number of cues such as epipolar consistency, limited range for disparity, photometric similarity, are now explored to reduce the number of potential edge correspondence. These cues are applied in sequence, each of which may be viewed as considered a weak classifier whose collective effect is significant, as in the cascade of classifiers in [23]. The result of applying the constraint posed by each cue, or cue filter, can be evaluated by measuring true positives (TP), false positives (FP), where an edge is considered veridical if within one pixel of a GT edge correspondence. We also define *ambiguity*, the number of potential edge correspondences that an edge forms beyond 1, Ambiguity $= \frac{TP+FP}{N} - 1$, where N is the number of edges that have stereo counterparts.

The Epipolar Proximity (EP) Filter: The corresponding edge to γ, $\overline{\gamma}$, is expected to lie on the epipolar line, *i.e.*, $\overline{\gamma}^\top [\mathcal{T}_0]_\times \mathcal{R}_0 \gamma = 0$. However, each edge has localization error, *e.g.*, on the order of 0.3 pixels [29] for the Third-Order

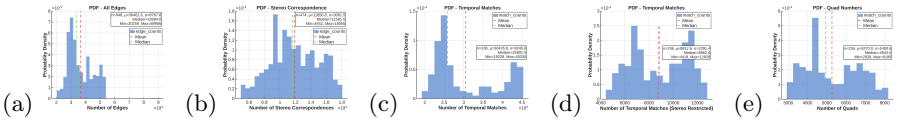


Figure 2: (a) Statistical distribution of edge populations across images in the ETH3D dataset [18]. Image size is 942×489. (b) The distribution of the number of stereo edge pairs across images. (c) The distribution of the number of temporal matches across frames. (d) The distribution of temporal matches restricted to edges with stereo edges correspondence. (e) The distribution of edge quadruplets.

Edge Detector [8]. Also, the pose $(\mathcal{R}_0, \mathcal{T}_0)$ has estimation error so that the corresponding edge is expected to have a small distance to the epipolar line. The distribution of this distance for veridical and non-veridical edge pairs is shown in Figure 3(a). Since the maximum veridical edge distance to the epipolar line is $\tau_{EP} = 0.5$ pixels, the EP filter removes all non-relevant correspondences with distance greater than τ_{EP}. The recall remains at 100% while precision improves to 3.6% from 0.005%, Table 1, leaving each edge with an average of 112 potential corresponding edges down from 48,631.

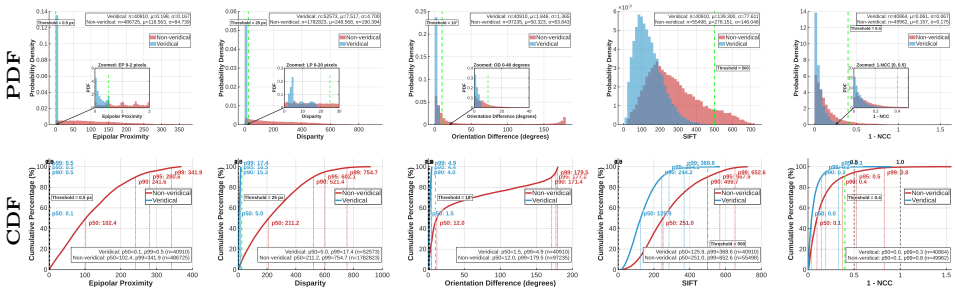
Limited Disparity (LD) Filter: Disparity $d = \gamma^+ - \gamma^-$, the discrepancy in position between corresponding points, is a function of depths, $\rho^\pm$ at $\gamma^\pm$,

$$d = \frac{\rho^- \mathcal{R}\gamma^- + \mathcal{T}}{(e_3^\top \mathcal{R}\gamma^-)\rho^- + e_3^\top \mathcal{T}} - \gamma^-. \tag{10}$$

A limited depth range implies a limited disparity range. Figure 3(b) shows that by thresholding disparity by τ_{LD} of, say, 23 pixels retains 100% recall, but raises precision to 52.4% from 3.6% with an ambiguity reduction from 112 to 7.2.

Orientation Similarity: The orientations of an edge in the left and right images, namely, $t^- = (\cos\theta^-, \sin\theta^-)$ and $t^+ = (\cos\theta^+, \sin\theta^+)$, respectively, is similar as measured by $\Delta\theta = \cos^{-1}({t^+}^\top t^-)$. Figure 3(c) shows that 99% of edges have orientation similarity under 5°. An application of this filter with $\tau_\theta = 5°$ maintains 100% recall, raises precision from 52.4% to 67.5%, and reduces average ambiguity from 7.2 to 4.9.

Photometric Similarity (NCC and SIFT): Each edge separates a left and a right neighborhood patch that reflects surface properties of the scene, Figure 4. It is expected that under small view perturbations, such as those under stereo and temporal viewing, these patches remain similar. Two measures of similarity, namely, NCC and SIFT are used to address both smooth and textured patches. Some edges, *e.g.*, reflectance edges, are expected to have similar patches on both sides of the edge, Figure 4. Let the patches to left and right of an edge be labeled as P_L and P_R. Then, for an edge correspondence $((\gamma^-, t^-), (\gamma^+, t^+))$ to be photometrically consistent, the paired patches (P_L^-, P_L^+) as well as (P_R^-, P_R^+) must



(a) **EP** (b) **LD** (c) **Orientation** (d) **SIFT** (e) **1-NCC**

Figure 3: The Probability Density Function (PDF) and the Cumulative Distribution Function (CDF) for different cue filters applied to veridical (shown in blue) and non-veridical (shown in red). (a) Epipolar proximity, $\tau_{EP} = 0.5$, (b) Limited disparity, $\tau_{LD} = 25$, (c) Orientation, $\tau_o = 10$, (d)SIFT, $\tau_{SIFT} = 500$, (e)NCC, $\tau_{NCC} = 0.6$.

Order	Filter name	Recall (%)	Precision (%)	Ambiguity Avg.	Ambiguity Max
0	Baseline	100	0.005	48631	48631
1	Epipolar Proximity	100	3.6	111.7	471
2	Limited Disparity	100	52.4	7.2	48
3	Orientation Similarity	100	67.5	4.9	46
4	NCC	100	70.8	4.5	46
5	SIFT	99.7	74.3	4.3	45
6	Best & Nearly Best (NCC)	99.6	81.5	3.5	45
7	Best & Nearly Best (SIFT)	99.5	86.6	2.7	45
8	Photo. Refin./Edge Clust.	98.4	96.3	0.1	15
9	NCC	98.3	96.4	0.093	1
10	Best	98	98	0	0

Table 1: Evaluations of applying cue filters to the stereo edge correspondences of the ETH3D [18] dataset.

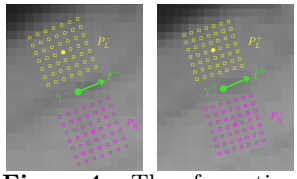


Figure 4: The formation of patches, (P_L^-, P_R^-) and (P_L^+, P_R^+), located at the two sides of γ^- and γ^+, respectively, enables robust NCC and SIFT similarity measurement between the two edges.

both be similar. However, for some edges, such as occluding edges, only one of the two paired patches need to be similar, *i.e.*, either (P_L^-, P_L^+) or (P_R^-, P_R^+). Furthermore, due to edge direction ambiguity, the correct pairing may be (P_L^-, P_R^+) and (P_R^-, P_L^+). Thus, $\max(\mu(P_L^-, P_L^+), \mu(P_R^-, P_R^+), \mu(P_L^-, P_R^+), \mu(P_R^-, P_L^+))$ is used, where μ is the photometric similarity measure, either based on NCC or SIFT. Figure 3(d) shows the distribution of the NCC similarity measure in the form of 1-NCC and its cumulative graph. First, the SIFT measure, Figure 3(e)., at 500 maintains 100% recall but raises precision from 67.5% to 70.8%, and reduces ambiguity from 4.9 to 4.5. Second, thresholding the NCC similarity at 0.6 loses a minute 0.3% in recall but raises the precision significantly from 70.8% to 74.3%, and reduces ambiguity from 4.5 to 4.3, Table 1.

Best and Nearly Best Filter: At this stage, each edge has about 5 potential correspondences. Selecting the most similar pair is brittle since the veridical edge may be a close second or third. Rather than selecting an optimal correspondence, the Best and Nearly Best (BNB) filter simply discard those that are not the best match nor do they come close to the best match. For the NCC we discard all matches below 90% of the best NCC, and for SIFT, discard those below 40% of the best SIFT similarity. The application of these two BNB filters reduces recall slightly from 99.7 to 99.6 and then to 99.5 while increasing precision from 74.3 to 81.5 and then to 86.5, reducing ambiguity from 4.3 to 3.5 and then to 2.7.

Photometric Refinement and Edge Clustering: Third-order edges detector produces a dense collection of 1-3 edges per pixel [29]. These edges share high similarities in visual cues, creating a hard condition for making a filtering decision. Observe, however, that these multiple edge hypotheses arise because a local curve is sampled densely and thus all hypotheses essentially represent the same curve. As such, they can be consolidated by creating an equivalent class of hypotheses. Locally, a curve can be approximated by a line. Thus, each edge is considered equivalent to that edge shifted along its tangent direction to the epipolar line that are supposed to lie on, a technique proposed by [27], without altering the underlying hypothesis.

Of course, this does not necessarily corrects the inaccuracies transversal to the curve remain. Thus, *photometric refinement* is applied to each edge hy-

pothesis, optimizing their locations along the epipolar line by minimizing their photometric residual. Specifically, photometric residual of an edge correspondence is defined by the NCC of patches identified by the maximum over the four pairs, *i.e.*, $(P_L^-, P_L^+), (P_R^-, P_R^+), (P_L^-, P_R^+), (P_R^-, P_L^+)$, after applying the photometric similarity. Through Gauss-Newton optimization, each edge hypothesis gets accurate location such that non-veridical edges that were originally close to the ground-truth are now turned to be veridical. Since no edge hypothesis is ruled out, ambiguity remains the same, but the precision increases significantly from 86.6% to 96.6% with a 1.1% drop in recall.

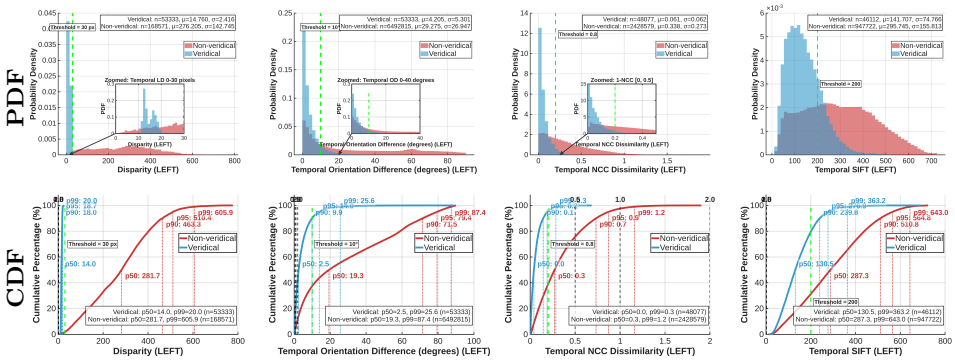
Refining the locations of multiple edge hypotheses in a close neighborhood brings them together within a pixel, while groups of edges are fairly distinct. Edges that are spatially close to each other now become redundant, and thus they can be consolidated by agglomerative clustering. For each cluster, one single representative edge is formed where the location is the cluster centroid while the orientation is Gaussian weighted averaged by all edges [27]. This drastically reduces the ambiguity, from 2.7 down to a mere 0.1, Table 1 while maintaining a stable recall of 98.4% with a negligible 0.3% drop for precision.

Best Filter: Edge hypothesis as a representation of a cluster of edges are significantly distinct so that photometric similarity can be safely applied again. Finally, a "Best" filter is adopted to pick the edge hypothesis whose NCC similarity is the highest. This enforces a strict one-to-one mapping between stereo edges, enabling unique matches serving as a preparatory step for simplifying the subsequent temporal matching across stereo frames. In the end, constructed stereo edge pairs arrives at around 98% recall and 98% precision with zero ambiguity.

5.2 Constructing Temporal Edge Correspondences to form Quads

The matching of edges across a temporal sequence is identical to the stereo case with two significant differences: *(i)* the relative pose between cameras over time is not available so that the EP filter cannot be applied, *(ii)* optimizing edge locations through minimizing photometric refinement is not constrained on the epipolar line and can move freely in the local neighborhood of the image domain, and *(iii)* the cameras typically move further between two adjacent frames than between left and right cameras in the stereo frames, implying that thresholds must be more generous. However, observe that for VO, producing highly accurate edge correspondences, namely, high precision, is more critical than maintaining a large quantity of edge correspondences, *i.e.*, high recall. Thus, we strict the thresholds for each filter, Figure 5 so that temporal edge pairs from the left and right sides achieve 83.56% and 81.78% precision. Finally, enforcing quadruplet consistency, the precision arrives at around 90% precision.

Forming Edge Quadruplet: The stereo and temporal matching of edges can now form a basis for forming quads. Stereo edge correspondences gives $((\gamma^-, t^-), (\gamma^+, t^+))$ pairs, while temporal edge correspondence gives $((\gamma^-, t^-), (\overline{\gamma}^-, \overline{t}^-))$ and $((\gamma^+, t^+), (\overline{\gamma}^+, \overline{t}^+))$. The formation of a quad requires that $(\overline{\gamma}^-, \overline{t}^-)$ and $(\overline{\gamma}^+, \overline{t}^+)$ forms a stereo edge pair. This constraint further improves the left and right sequential correspondences as the last line of Table 2 shows. The quads



(a) **LD** (b) **Orientation** (c) **1-NCC** (d) **SIFT**

Figure 5: The Probability Density Function (PDF) and the Cumulative Distribution Function (CDF) for different cue filters applying to the temporal edge correspondences.

that are formed as a result of this process can now be compared against ground-truths, showing a 88.6% recall, 87.4% precision, and average ambiguity of 0.05, Figure 6 and Table 3.

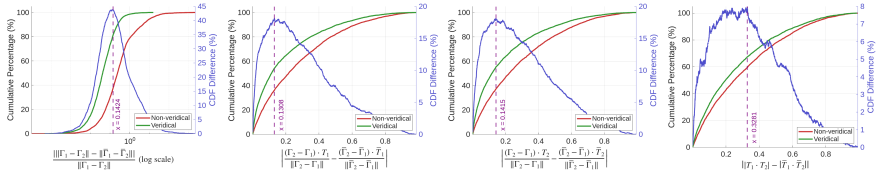


Figure 6: Cumulative probability of pairs of veridical and non-veridical edge quadruplets on which the constraints, Equation 9, apply.

6 QuadEdge-VO Framework

The formation of quads, described in Section 5, enables relative pose estimation from a minimal set of two quads, as presented in Section 4. In the presence of non-veridical quads, pose estimation is performed within a RANSAC framework. Specifically, in each RANSAC iteration, a relative pose hypothesis is generated

Order	Filter name	Left			Right		
		Recall (%)	Precision (%)	Ambiguity	Recall (%)	Precision(%)	Ambiguity
0	Baseline	100.00	0.01	16860.00	100.00	0.01	14148.00
1	Limited Disparity	100.00	1.28	463.53	100.00	1.26	472.98
2	Orientation Similarity	96.07	5.95	174.74	96.85	5.98	183.12
3	NCC	93.53	24.53	71.07	94.34	23.49	73.68
4	SIFT	82.66	48.38	26.13	82.52	46.86	27.08
5	Best & Nearly Best (NCC)	82.52	51.38	21.78	82.33	49.96	22.30
6	Best & Nearly Best (SIFT)	72.25	67.40	2.12	72.01	65.66	2.35
8	Photo. Refin./Edge Clust.	78.00	83.56	0.26	76.41	81.78	0.27
9	Quadruplet Consistency	50.62	90.79	0.05	50.01	89.56	0.05

Table 2: Evaluations on the temporal edge correspondences between the images of the two consecutive stereo rigs of the entire ETH3D [18] dataset.

Order	Constraint	Recall (%)	Precision (%)
0	Baseline	100.00	87.04
1	$\frac{\|\Gamma_2-\Gamma_1\|-\|\overline{\Gamma}_2-\overline{\Gamma}_1\|}{\|\Gamma_2-\Gamma_1\|}$	80.73	89.77
2	$\left\|\frac{(\Gamma_2-\Gamma_1)\cdot T_1}{\|\Gamma_2-\Gamma_1\|}-\frac{(\overline{\Gamma}_2-\overline{\Gamma}_1)\cdot\overline{T}_1}{\|\overline{\Gamma}_2-\overline{\Gamma}_1\|}\right\|$	42.39	91.60
3	$\left\|\frac{(\Gamma_2-\Gamma_1)\cdot T_2}{\|\Gamma_2-\Gamma_1\|}-\frac{(\overline{\Gamma}_2-\overline{\Gamma}_1)\cdot\overline{T}_2}{\|\overline{\Gamma}_2-\overline{\Gamma}_1\|}\right\|$	22.16	93.49
4	$\left\|\|T_1\cdot T_2\|-\|\overline{T}_1\cdot\overline{T}_2\|\right\|$	19.40	94.67

Table 3: Evaluations of applying the constraints on a subset of quad pairs.

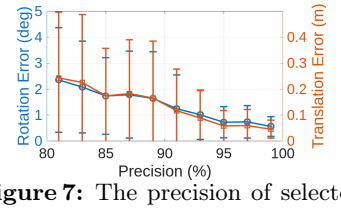


Figure 7: The precision of selected quad pairs satisfying the constraints are highly correlated with the relative pose estimation accuracy.

from a pair of selected quads that satisfies the constraints in Equation 9. The resulting pose hypothesis is then validated against all available quads. A quad is considered an inlier if both its reprojection error and orientation error are below 1.5 pixels and 10 degrees, respectively.

For a moving stereo rig, its motion trajectory can be recovered by estimating the relative pose with respect to a "key stereo rig", analogous to a keyframe used in typical visual odometry pipelines to reduce accumulated drift. The key stereo rig is selected based on heuristic but robust strategies [5], and remains active for a short temporal window, serving as the reference frame with respect to which the poses of subsequent stereo rigs are estimated. Observe, however, that temporal quads formed between stereo rigs several frames apart from the key stereo rig tend to be less reliable, as larger viewpoint changes are unfriendly for a reliable quad formation. We address this by first forming quads between adjacent stereo rigs, where the viewpoint change is small and correspondences are therefore more robust. Quads between temporally distant stereo rigs are then obtained indirectly by propagating correspondences through intermediate rigs. This enables reliable quad formation between rigs that are multiple frames apart without requiring direct matching across large temporal gaps.

In terms of efficiency, the third-order edge detector runs on a GPU requires less than 1 ms per image. The primary computational cost arises from constructing stereo and temporal edge pairs. When executed on a 32-core CPU with OpenMP parallelization, the full pipeline from third-order edge detection to relative pose estimation takes around 5 seconds per stereo frame pair. However, these operations are highly parallelizable, as individual edges can be processed independently. Moving this stage to the GPU thus offers a clear path toward real-time performance.

7 Experiments

Dataset and Competing Methods: The ETH3D [18] dataset is used to evaluate the performance of the proposed QuadEdge-VO. Specifically, it provides a variety of scenarios including indoor, outdoor, texture-rich and textureless scenes. In particular, there are scenes exhibiting curvilinear structures on a homogeneous surface. In this case, it is natural to examine whether methods relying on point features in conjunction with lines can perform effectively in this setting as no evident straight lines are present. Thus, recent point-line based approaches

such as Structure-PLP-SLAM [19], ORB-LINE-SLAM [1], are used as competing baselines. Additionally, MAC-VO [16], a modern learned point-based stereo odometry, is also included as a complementary method. Default settings are used for all three methods throughout the experiments. For MAC-VO, we use its "Performant Mode" to achieve the best performance.

Metrics: The performances are measured by estimation accuracy in terms of absolute pose error (APE) and relative pose error (RPE) [20]. Both metrics are decoupled into the rotation ($\text{APE}_{\mathcal{R}}$, $\text{RPE}_{\mathcal{R}}$) and the translation ($\text{APE}_{\mathcal{T}}$, $\text{RPE}_{\mathcal{T}}$) parts, in units of degrees and meters, respectively. Note that rather than applying full trajectory alignment with the ground truth, we only align the trajectory origins to better reflect real-world situations.

Qualitative Results: Figure 8 presents the constructed and veridical stereo edge correspondences, each represented by an individual but consistent color across the two views. Evidently, third-order edges are ample in the presence of texture and reflectance discontinuities, and the proposed stereo edge pairing through progressive filtering establishes rich (high recall) and precise (high precision) edge correspondences. On the other hand, among the constructed edge correspondences that are non-veridical, the majority of them are *inaccurate*: they are in the 1-10 pixels local neighborhood of the ground-truth edge location, Figure 9(a). Only a few non-veridical edge pairs are incorrect, Figure 9(b).

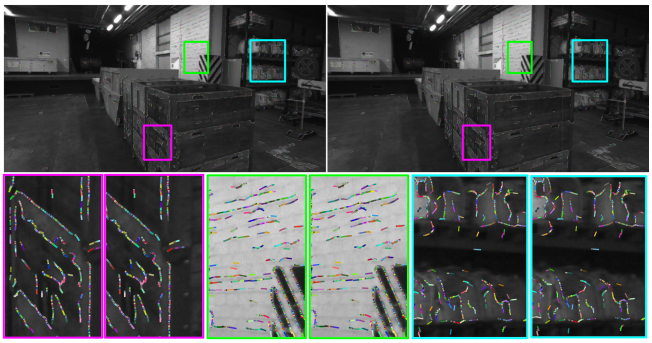


Figure 8: Correct stereo edge correspondences (16,039 edge pairs in this example), defined as in the 1 pixel vicinity of the ground-truth edge. The top image shows the full stereo image pair superimposed with edge correspondences in individual colors, while the bottom row provides detailed, zoomed-in views of three specific regions.

Figure 10 shows selected trajectories produced by QuadEdge-VO and the competing methods. Evidently, QuadEdge-VO recovers the geometry of the ground-truth trajectory much better than other approaches. Notably, in the

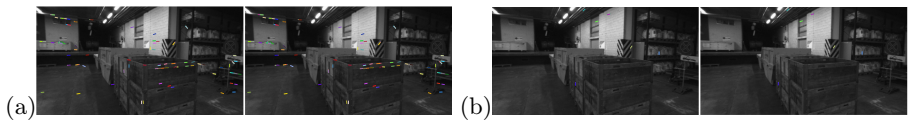


Figure 9: (a) Inaccurate stereo edge correspondences (242 edge pairs in this example) and (b) false stereo edge correspondences (11 edge pairs in this example) defined as 1-10 pixels and over 10 pixels away from the ground-truth edge, respectively.

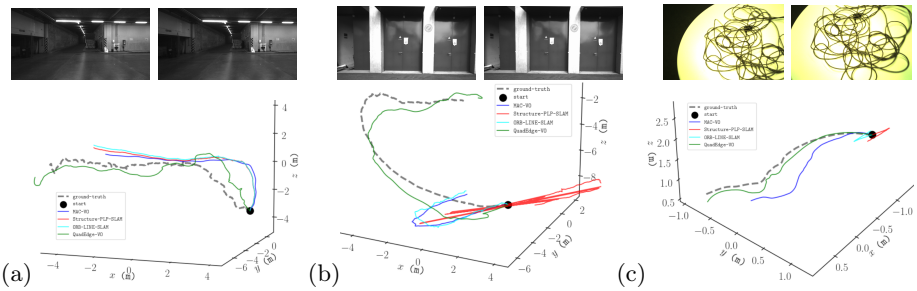


Figure 10: Qualitative results on the trajectories of Structure-PLP-SLAM [19], ORB-LINE-SLAM [1], MAC-VO [16], and the proposed QuadEdge-VO on the sequences of (a) delivery area, (b) electro, and (c) cable_2 of the ETH3D dataset, with one example stereo image pair shown on the top of each trajectory.

cable_2 sequence, both Structure-PLP-SLAM and ORB-LINE-SLAM failed to estimate good quality of camera poses, leading to no evident recovery of motion trajectory. We observed that this is due to very unreliable point feature correspondences, and even though lines are used, in the case of cluttered cables where no distinct line features present in the scene, the detected lines are too unstable to participate in the camera pose estimation process.

Quantitative Results: Table 4 quantitatively compares the average trajectory estimation accuracy of the proposed QuadEdge-VO and the other methods using the ETH3D dataset. Clearly, QuadEdge-VO provides outstanding performances, demonstrating the promising effectiveness of using edges for a stereo VO.

8 Conclusion

We presented QuadEdge-VO, a paradigm shift in edge-based odometry that replaces iterative minimization with a discrete, quad-based geometric solution. By pairing stereo and temporal edges into "edge quads", we reduce the 6-DoF pose estimation to a tractable constrained two-quad estimation. This approach is made computationally feasible by a cascaded filtering architecture that tames the combinatorial explosion of edge matches, achieving high precision without sacrificing subpixel accuracy. Experimental validation on the ETH3D dataset demonstrates that QuadEdge-VO significantly outperforms traditional methods in both accuracy and efficiency. Future work will focus on integrating these quad primitives into large-scale bundle adjustment for global SLAM consistency.

Methods	**Feature Type**	$\mathbf{APE}_{\mathcal{R}}$ (deg)		$\mathbf{APE}_{\mathcal{T}}$ (m)		$\mathbf{RPE}_{\mathcal{R}}$ (deg)		$\mathbf{RPE}_{\mathcal{T}}$ (m)	
		mean	std	mean	std	mean	std	mean	std
MAC-VO [16]	Points	1.24	0.71	0.97	0.42	0.020	0.027	0.018	**0.021**
Structure-PLP-SLAM [19]	Points+Lines	2.04	0.88	1.37	0.59	0.023	0.031	0.051	0.095
ORB-LINE-SLAM [1]	Points+Lines	1.59	0.84	1.27	0.61	0.051	0.046	0.052	0.103
QuadEdge-VO	Edges	**0.07**	**0.03**	**0.82**	**0.38**	**0.011**	**0.018**	**0.017**	**0.021**

Table 4: Quantitative evaluations of the proposed QuadEdge-VO against the competing methods on the ETH3D dataset. **Bold**: the best.

References

1. Alamanos, I., Tzafestas, C.: ORB-LINE-SLAM: an open-source stereo visual SLAM system with point and line features. Authorea Preprints (2023)
2. Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J.: Mocha-stereo: Motif channel attention network for stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 27768–27777 (2024)
3. Christensen, K., Hebert, M.: Edge-direct visual odometry. arXiv preprint arXiv:1906.04838 (2019)
4. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence **40**(3), 611–625 (2017)
5. Fanfani, M., Bellavia, F., Colombo, C.: Accurate keyframe selection and keypoint tracking for robust visual odometry. Machine Vision and Applications **27**(6), 833–844 (2016)
6. Huang, H., Li, L., Cheng, H., Yeung, S.K.: Photo-SLAM: Real-time simultaneous localization and photorealistic mapping for monocular stereo and RGB-D cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21584–21593 (2024)
7. Kim, C., Kim, J., Kim, H.J.: Edge-based visual odometry with stereo cameras using multiple oriented quadtrees. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5917–5924. IEEE (2020)
8. Kimia, B.B., Li, X., Guo, Y., Tamrakar, A.: Differential geometry in edge detection: accurate estimation of position, orientation and curvature. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1573–1586 (2018)
9. Kuse, M., Shen, S.: Robust camera motion estimation using direct edge alignment and sub-gradient method. In: 2016 IEEE international conference on robotics and automation (ICRA). pp. 573–579. IEEE (2016)
10. Liu, C., Yu, H., Cheng, P., Sun, W., Civera, J., Chen, X.: PE-VINS: Accurate monocular visual-inertial SLAM with point-edge features. IEEE Transactions on Intelligent Vehicles (2024)
11. Liu, M., Zuo, X., Huang, R., Zhao, M., Chen, J., Li, L.: ROEVO: Robust organized edge feature-based visual odometry using RGB-D cameras. IEEE Transactions on Robotics (2025)
12. Lu, X., Du, S., Yan, Y., Lu, X., Ikenaga, T.: Towards free-form local feature matching. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
13. Maity, S., Saha, A., Bhowmick, B.: Edge SLAM: Edge points based monocular visual slam. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2408–2417 (2017)
14. Mo, J., Islam, M.J., Sattar, J.: Fast direct stereo visual SLAM. IEEE Robotics and Automation Letters **7**(2), 778–785 (2021)
15. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras. IEEE transactions on robotics **33**(5), 1255–1262 (2017)
16. Qiu, Y., Chen, Y., Zhang, Z., Wang, W., Scherer, S.: MAC-VO: Metrics-aware covariance for learning-based stereo visual odometry mac-vo. github. io. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 3803–3814. IEEE (2025)
17. Schenk, F., Fraundorfer, F.: RESLAM: A real-time robust edge-based SLAM system. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 154–160. IEEE (2019)

18. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3260–3269 (2017)
19. Shu, F., Wang, J., Pagani, A., Stricker, D.: Structure plp-slam: Efficient sparse mapping and localization using point, line and plane for monocular, rgb-d and stereo cameras. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2105–2112. IEEE (2023)
20. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 573–580. IEEE (2012)
21. Tarrio, J.J., Pedre, S.: Realtime edge-based visual odometry for a monocular camera. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 702–710 (2015)
22. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. Advances in Neural Information Processing Systems **36**, 39033–39051 (2023)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. vol. 1, pp. I–I. Ieee (2001)
24. Wang, X., Xu, G., Jia, H., Yang, X.: Selective-stereo: Adaptive frequency information selection for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19701–19710 (2024)
25. Wang, X., Dong, W., Zhou, M., Li, R., Zha, H.: Edge enhanced direct visual odometry. In: BMVC (2016)
26. Yin, H., Liu, P.X., Zheng, M.: Ego-motion estimation with stereo cameras using efficient 3D-2D edge correspondences. IEEE Transactions on Instrumentation and Measurement **71**, 1–11 (2022)
27. Zhang, Q., Chien, C.H., Fabbri, R., Kimia, B.: 3D Curvix: From multiview 2D edges to 3D curve segments. In: The 36th British Machine Vision Conference (BMVC) (2025)
28. Zhao, H., Shang, J., Liu, K., Chen, C., Gu, F.: EdgeVO: An efficient and accurate edge-based visual odometry. arXiv preprint arXiv:2302.09493 (2023)
29. Zheng, Y., Chien, C.H., Fabbri, R., Kimia, B.: 3D edge sketch from multiview images. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3196–3205. IEEE (2025)
30. Zhou, Y., Li, H., Kneip, L.: Canny-VO: Visual odometry with RGB-D cameras based on geometric 3D-2D edge alignment. IEEE Transactions on Robotics **35**(1), 184–199 (2018)