2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Robust Monocular Edge Visual Odometry through Coarse-to-Fine Data Association

Xiaolong Wu, Patricio A. Vela, and Cédric Pradalier

***Abstract*— This work describes a monocular visual odometry framework, which exploits the best attributes of edge features for illumination-robust camera tracking, while at the same time ameliorating the performance degradation of edge mapping. In the front-end, an ICP-based edge registration provides robust motion estimation and coarse data association under lighting changes. In the back-end, a novel edge-guided data association pipeline searches for the best photometrically matched points along geometrically possible edges through template matching, so that the matches can be further refined in later bundle adjustment. The core of our proposed data association strategy lies in a point-to-edge geometric uncertainty analysis, which analytically derives (1) a probabilistic search length formula that significantly reduces the search space and (2) a geometric confidence metric for mapping degradation detection based on the predicted depth uncertainty. Moreover, a match confidence based patch size adaption strategy is integrated into our pipeline to reduce matching ambiguity. We present extensive analysis and evaluation of our proposed system on synthetic and real-world benchmark datasets under the influence of illumination changes and large camera motions, where our proposed system outperforms current state-of-art algorithms.**

I. INTRODUCTION

In recent decades, monocular Visual Odometry (VO) systems have shown their full potential to assist various outdoor robotic applications. Among these algorithms, indirect methods [1] are *de facto* standards due to the robustness of visual features against both photometric noise and lens distortion, while direct methods present [2] better motion estimation robustness contributed by its more complete usage of information contained in the image. Point features, widely used by both approaches, are known to fail under specific conditions such as sudden lighting changes, large camera motions, or texture-less environments, which is attributed to low feature detections or associations across frames.

Edges are alternatives to point features with improved robustness to the aforementioned situations. They are geometric features extracted from raw images with inter-frame edge registration performed using iterative-closest-point (ICP) based direct alignment [3] [4] [5]. The first edge VO [6] aligned edges by searching for the closest counterpart along the normal direction. Later, efficiency improvements for 2D-3D registration based on the Distance Transform (DT) [3] improved the real-time properties of motion estimation. The optimizability of this formula is further improved by substituting DT with Approximate Nearest Neighbor Fields (ANNFs) [4] or Oriented Nearest Neighbor Fields (ONNFs) [5], which have demonstrated strong performances for RGB-D sensors. For RGB-D sensors these techniques work well. Meanwhile, image-only monocular edge-based VO remains challenging since binary edge features lead to partially observable depth under ICP-based mapping frameworks. Conventional depth estimation algorithms that search for best matches along epipolar lines may result in unreliable correspondences and lead to unstable and error-prone depth estimates in Fig. 1.

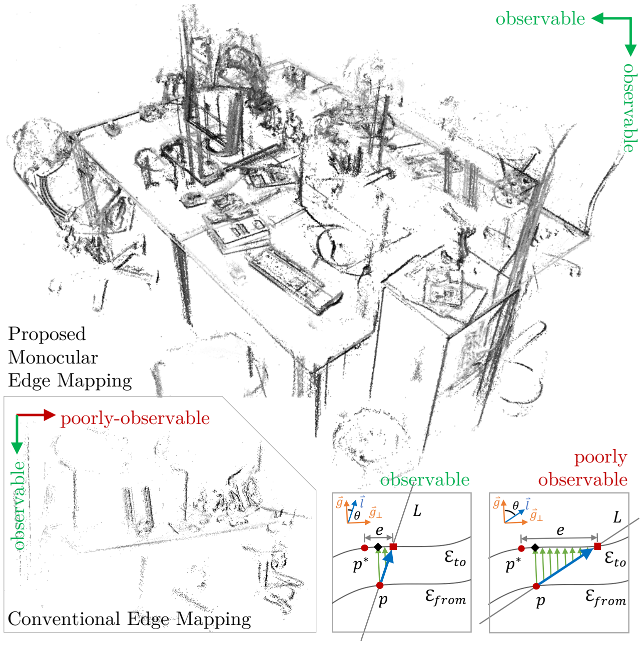

Fig. 1: **Our proposed Monocular Edge VO** (top) is capable of using edge feature and image gradient to overcome the partial observability issue of pure Edge Mapping (bottom-left). Pure edge mapping minimize point-to-tangent error (↑) that results in more erroneous matches (🟥) under poorly-observable direction (bottom-right) than observable direction (bottom-middle).

To address the partial observability issue of edge mapping, more sophisticated group matching strategies can realize geometrically consistent matches [7], but fail to provide any theoretical guarantees on correctness. Using optical flow [8] to find the matches through photometric minimization makes the depth fully observable in the back-end [9]. However, most optical flow methods rely on the brightness constancy assumption, limiting the set of situations for which edges complement point features.

[1] Xiaolong Wu and Patricio A. Vela are with the School of Electrical and Computer Engineering, Atlanta, GA 30332, United States xwu,pvela@gatech.edu

[2] Cédric Pradalier with the School of Interactive Computing, UMI2958 GeorgiaTech-CNRS, Metz 57070, France cedric.pradalier@georgiatech-metz.fr

*The work was supported in part by NSF Award #1816138.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

As an alternative, illumination-robust tracking algorithms based on the Lucas-Kanade method [10] have also been studied extensively [11]. Analyses suggest that gradient [12] and census transform [13] approaches show state-of-the-art tracking accuracy, but considerably compromise the convergence basin. The reduced convergence radius arises from a flatter cost functions, which increases the sensitivity to perception noise and introduces artificial local optima.

One way to have more illumination-robust template matching is to utilize confidence measures [14] to assess the correctness of hypotheses; a technique used in the field of stereo vision for error detection [15]. Of the various metrics, Attainable Maximum Likelihood (AML) [16] shows high performance for multi-view stereo matching, which is readily incorporated into our system for match ambiguity detection.

Inspired by semi-direct methods [17] [18] integrating the complementary strengths of indirect and direct formulations for superior performance, we propose a monocular edge VO framework in Fig. 2 integrating the illumination-robustness of edge features, the informativeness of photometric matching, and the efficiency of pose-graph optimization to solve the issues mentioned above. Our proposed framework inherently conforms to a coarse-to-fine data association structure, which iteratively refines the edge point correspondences by exploiting geometric and photometric information. The main contributions of this work ares:

- A monocular edge VO framework, comprised of ICP-based edge alignment, edge-guided data association, and local BA, which is capable of performing illumination-robust camera tracking and scene reconstruction without incurring edge mapping degradation.
- An edge-guided data association pipeline incorporating probabilistic search length approximation, image-gradient-based template matching, match-confidence-based patch size adaption, and depth-confidence-based match conditioning.
- A point-to-edge geometric uncertainty analysis that analytically derives a probabilistic search length formula and a depth confidence measure that improves the efficiency and accuracy of our proposed system.

II. SYSTEM OVERVIEW

Fig. 2 illustrates our proposed monocular edge VO framework, which consists of two parallel threads named edge tracking and mapping. Our proposed edge tracking threads aim to coarsely estimate the camera motion and edge point association from the current frame to the latest keyframe, which is achieved through the minimization of point-to-tangent error in Eq. (2a) using an ICP pipeline. As soon as a new keyframe is created, the edge-guided data association module in Sec. IV refines the correspondences by incorporating illumination-robust photometric information. Finally, the local BA jointly optimizes the camera poses and scene structure using resultant matches in Eq. (2b) within a sliding window of keyframes. We also provide the option to involve the image-gradient-based costs Eq. (2c) into both edge tracking and local BA layers. The hybrid option is designed for applications that prioritize accuracy over high-speed operation.

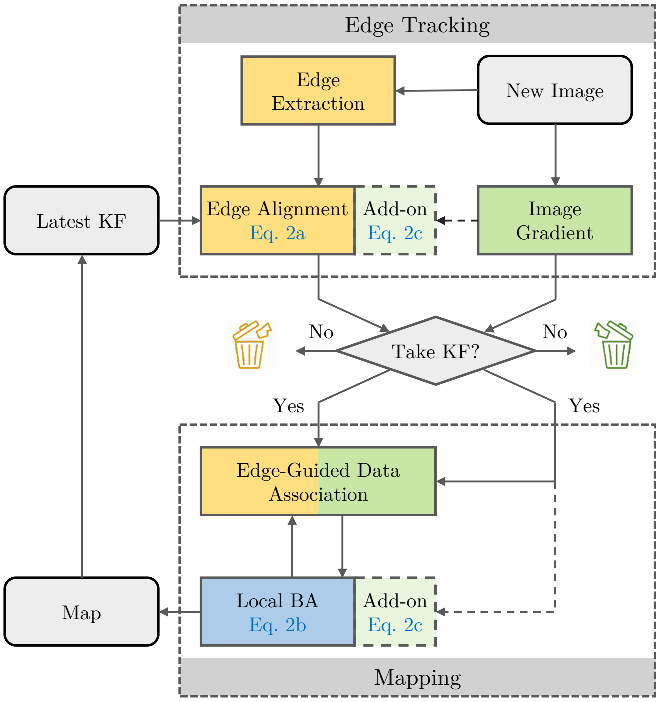

Fig. 2: **Our proposed monocular Edge VO framework** is a KeyFrame (KF) based monocular VO framework, which can be generally divided into edge alignment, edge-guided data association, and local BA.

Note that the edge tracking, local BA, and keyframe selection are well-studied problems, therefore we choose the state-of-the-art implementations for our system design. Our edge alignment front-end implements the ANNFs [4] [5] with a pyramidal coarse-to-fine scheme for point-to-tangent registration in Eq. (2a), which approximates the nearest neighbors as temporal correspondences for ICP-based optimization. Our local bundle adjustment algorithm performs a joint optimization of camera poses and scene structure altogether in Eq. (2b), which can be achieved through pose-graph optimization [19] [20] or a customized solver [2]. Besides, we implement the distance-based keyframe selection strategy that has been successfully demonstrated in DSO [2].

III. PROBLEM FORMULATION

Consider an acquired image in the reference frame $I_r : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^2$ is the image domain. A 3D scene point $\mathbf{P} = (x,y,z)^T$ is parameterized by its inverse depth $d = z^{-1}$. Each pixel $\mathbf{p} = (u,v)^T \in \Omega$ can be back-projected into the 3D world using the back-projection function $\mathbf{P} = \pi^{-1}(\mathbf{p})$, and inversely using projective warp function $\mathbf{p} = \pi(\mathbf{P})$. A 3D rigid body transformation $\mathbf{G} \in SE(3)$ consists of a rotation $\mathbf{R} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$. In the optimization framework, $\mathbf{G}$ is represented as its corresponding Lie Group component $\xi \in \mathfrak{se}(3)$, where this element can be mapped to $\mathbf{G} \in SE(3)$ through the exponential mapping described in [21].

Given an arbitrary pixel selector $\mathcal{S}(\cdot)$, the group of edge pixels $\mathcal{S}_r = \{\mathbf{p}_r\}$ are subsequently projected to current frame

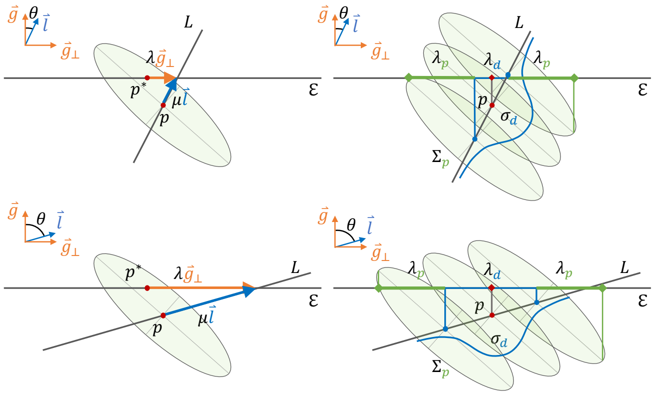

Fig. 3: **The Point-to-Edge Uncertainty Analysis** for observable (top: θ is small) and poorly-observable (bottom: θ is large) cases. The geometric relationship described in Eq. (3b) (left) and the geometric interpretation of our derived edge search length in Eq. (3a) (right) are presented.

k as:

$$\mathbf{p}_{kr} = \pi(\mathbf{R}_{kr}\pi^{-1}(\mathbf{p}_r, d_r) + \mathbf{t}_{kr}) \tag{1}$$

The point-to-tangent edge alignment error $E_{kr}^{\mathcal{E}}$, reprojection error $E_{kr}^{\mathcal{R}}$, and generalized photometric error $E_{kr}^{\mathcal{P}}$ from the reference frame to frame k can be generally written as:

$$E_{kr}^{\mathcal{E}} := \sum_{\mathbf{p}_r \in \mathcal{S}_r^{\mathcal{E}}} w_{\mathbf{p}_r}^{\mathcal{E}} \|\mathbf{g}^T(\mathbf{p}_{kr} - n(\mathbf{p}_{kr}))\|_\gamma \tag{2a}$$

$$E_{kr}^{\mathcal{R}} := \sum_{\mathbf{p}_r \in \mathcal{S}_r^{\mathcal{R}}} w_{\mathbf{p}_r}^{\mathcal{R}} \|\mathbf{p}_{kr} - \mathbf{p}_k\|_\gamma \tag{2b}$$

$$E_{kr}^{\mathcal{P}} := \sum_{\mathbf{p}_r \in \mathcal{S}_r^{\mathcal{P}}} w_{\mathbf{p}_r}^{\mathcal{P}} \|F_k(\mathbf{p}_{kr}) - F_r(\mathbf{p}_r)\|_\gamma \tag{2c}$$

where $w_{\mathbf{p}_r}$ is the weight assigned for each selected pixel from $\mathcal{S}_r$ in the reference frame, and $\|\cdot\|_\gamma$ is the Huber norm. Specifically for each formulation, $n(\cdot)$ represents the nearest neighbor edge pixel in current frame k using the Euclidean distance metric, $\mathbf{g}$ is the abbreviation of $\mathbf{g}(n(\mathbf{p}_{kr}))$ representing the gradient direction vector of the temporal match of a given projected pixel $\mathbf{p}_{kr}$, $\{\mathbf{p}_r, \mathbf{p}_k\}$ is a pair of matched points between the reference image and image k, and $F(\cdot)$ represents any representation calculated from image I, such as intensity or gradient.

IV. EDGE-GUIDED DATA ASSOCIATION

This section presents a point-to-edge uncertainty analysis (Sec. IV-A). The analysis informs our edge-guided data association pipeline (Sec. IV-B), which further refines the edge-point correspondences using geometric relationships and photometric information.

A. Point-to-Edge Uncertainty Analysis

Point-to-edge analysis is carried out to realize the potential search length, that is, the error variance of the search radius along the edge direction caused by tracking error on ξ and inverse depth error on d. To simplify the analysis, we make two assumptions: (1) the edge is locally linear so that its gradient direction is locally constant, and (2) the rotation error on ξ plays a minor role in the epipolar line direction estimation. After such simplifications, the search line along edge direction $\mathcal{S}$ and epipolar line $\mathcal{L}$ are approximated as:

$$\mathcal{S} := \{\mathbf{p}^* + \lambda\mathbf{g}_\perp\} \quad \mathcal{L} := \{\mathbf{p} + \mu\mathbf{l}\} \tag{3a}$$

$$\mathbf{p}^* + \lambda\mathbf{g}_\perp = \mathbf{p} + \mu\mathbf{l} \tag{3b}$$

where $\mathbf{p}$ represents the reprojected edge point from the reference frame to the current frame, and $\mathbf{p}^*$ denotes its correspondence lying on the locally linear region of target edge. $\mathbf{g}_\perp$ and $\mathbf{l}$ are the normalized perpendicular epipolar line and image gradient directions, while λ and μ are their distance factors. θ represents the angle between $\mathbf{g}$ and $\mathbf{l}$, which describe the angular relationship between $\mathcal{S}$ and $\mathcal{L}$. The described geometric relationship is illustrated in Fig. 3.

1) Probabilistic search length: Here, we derive the formula for the search length factor λ and its variance σ_λ^2 w.r.t. the μ and $\mathbf{p}$. For simplicity, we assume the uncertainties of μ and $\mathbf{p}$ are independent, so that the derived search length function λ and its variance σ_λ^2 can be expressed as:

$$\lambda(\mathbf{p},\mu) = \langle\mathbf{p}^* - \mathbf{p}, \mathbf{g}_\perp\rangle + \mu\langle\mathbf{l}, \mathbf{g}_\perp\rangle = e_{p\perp g} + \mu sin\theta \tag{4a}$$

$$\sigma_\lambda^2(\mathbf{p},\mu) = \mathbf{J}_p\Sigma_p\mathbf{J}_p^T + \mathbf{J}_\mu\sigma_\mu^2\mathbf{J}_\mu^T = \sigma_{p\perp g}^2 + \sigma_\mu^2\sin^2\theta \tag{4b}$$

where $\mathbf{J}_p$ and $\mathbf{J}_\mu$ are the Jacobians of Eq. (4a) given the statistics of $\mathbf{p}$ and μ. $e_{p\perp g}$ is the reprojection error component perpendicular to the edge normal direction $\mathbf{g}$. $\Sigma_\mathbf{p}$ and σ_μ^2 denote the (co-)variances of reprojected point and depth disparity. For fast calculation, the upper bound of variance can be estimated using the inequality relationship as:

$$\sigma_\lambda = \|\sigma_{p\perp g}^2 + \sigma_\mu^2\sin^2\theta\|^{\frac{1}{2}} \leq \sigma_{p\perp g} + \sigma_\mu|\sin\theta| \tag{5}$$

Setting the center of search to be the temporally estimated edge point correspondence $n(\mathbf{p})$, the search radius $\lambda_{1/2}$ can be expressed as:

$$\lambda_{1/2} = k_p\sigma_{p\perp g} + k_\mu\sigma_\mu|\sin\theta| \tag{6}$$

where k_p and k_μ denote the gains to compensate the potential shrinkage of search length due to our approximations.

The point reprojection covariance Σ_p can be readily calculated from edge alignment front-end using conventional uncertainty propagation, which can be decomposed into a group of more compact representations, the eigenvalue σ and eigenvector $\mathbf{v}$, through eigenvalue decomposition as follows:

$$\Sigma_p = \mathbf{J}_\xi^p\left(\sum_p \mathbf{J}_\xi^{r\,T}\mathbf{J}_\xi^r\right)^{-1}\mathbf{J}_\xi^{p\,T}\sigma_r^2 = \begin{bmatrix}\mathbf{v}_1\mathbf{v}_2\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{bmatrix}\begin{bmatrix}\mathbf{v}_1^T\\ \mathbf{v}_2^T\end{bmatrix} \tag{7}$$

where σ_r^2 and $\mathbf{J}_\xi^r$ represent the variance of residual and its individual Jacobian vector of point $\mathbf{p}$ w.r.t. camera pose ξ in Eq. (2a), and $\mathbf{J}_\xi^p$ denotes the Jacobian in Eq. (1). Therefore, the two components of search length representation can be upper bounded by enforcing the symmetric structure of search length as follows:

$$\sigma_{p\perp g} = \max\left(\sigma_1\langle\mathbf{v}_1, \mathbf{g}_\perp\rangle, \sigma_2\langle\mathbf{v}_2, \mathbf{g}_\perp\rangle\right) \tag{8a}$$

$$\sigma_\mu = \max\left(\|\mathbf{p}(\xi, d \pm \sigma_d) - \mathbf{p}(\xi, d)\|_2\right) \tag{8b}$$

where $\mathbf{p}(\xi, d)$ denotes the point reprojection function described in Eq. (1). The geometric interpretation of the derived search length centered at temporal correspondence (♦) is illustrated in Fig. 3.

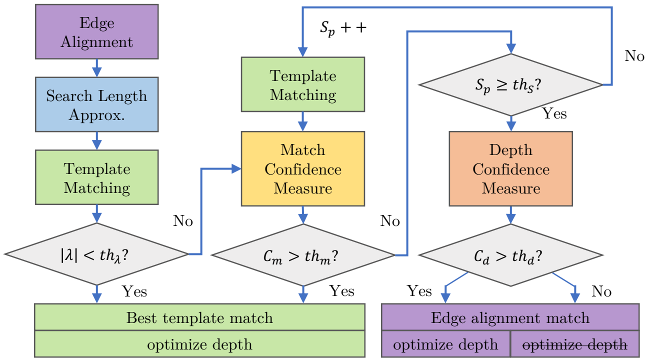

Fig. 4: **Edge-Guided Data Association Pipeline** incorporates probabilistic search length approximation (Sec. IV-B.1), illumination-robust template matching (Sec. IV-B.2) with match confidence based patch size adaption (Sec. IV-B.3), and depth confidence based conditioning (Sec. IV-B.4). Note that the edge alignment match comes directly from the proposed front-end, which doesn't involve any calculation here.

2) Depth uncertainty: Eliminating the search length factor λ from Eq. (3b), the disparity estimate μ and its variance σ_μ^2 can be expressed as:

$$\mu(\mathbf{p}) = \frac{\langle \mathbf{p}^* - \mathbf{p}, \mathbf{g} \rangle}{\langle \mathbf{g}, \mathbf{l} \rangle} = \frac{e_{p\|g}}{cos\theta} \tag{9a}$$

$$\sigma_\mu^2 = \mathbf{J}_p \Sigma_p \mathbf{J}_p^T = \frac{\sigma_{p\|g}^2}{\cos^2\theta}, \tag{9b}$$

where $\mathbf{J}_p$ is the Jacobian of Eqn. (9a) given the statistics of $\mathbf{p}$. $e_{p\|g}$ is the reprojection error component that is parallel to edge normal direction $\mathbf{g}$. Similar to the derivation in Eq. (8a), $\sigma_{p\|g}$ can be expressed as follows:

$$\sigma_{p\|g} = \max\left(\sigma_1 \langle \mathbf{v}_1, \mathbf{g} \rangle, \sigma_2 \langle \mathbf{v}_2, \mathbf{g} \rangle\right). \tag{10}$$

B. Edge-Guided Data Association Pipeline

The heart of our solution involves the proposed edge-guided data association pipeline, whose processing flow is depicted in Figure 4 with input of edge alignment matches 🟪. The important components include search length approximation 🟦, illumination-robust template matching 🟩, match-confidence-based patch size adaption 🟨, and depth confidence based match conditioning 🟧.

1) Search length approximation: Given the camera transformation from the edge alignment front-end, all edge points can be projected onto newly added keyframe. Their nearest neighbors in a new keyframe then serve as the coarse initialization of edge point correspondence for further refinement in Fig. 5 (1). Then, the search length λ is calculated for each point of interest using Eq. (6) based on their statistics.

2) Illumination-robust template matching: Given the estimated search radius, our proposed probabilistic 1D search strategy starts with the coarsely estimated edge point correspondence. A standard region growing algorithm is used to explore the nearby edge points for template matching. To compensate for the rotation and scale difference between the matched patches, the patch is pre-transformed in Fig. 5 (2) based on estimated parameters $\mathbf{R}$, $\mathbf{t}$, and $\mathbf{d}$ by making the assumption that the transformation is locally constant.

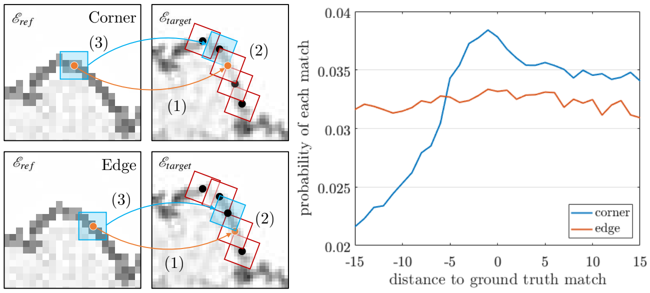

Fig. 5: **Our proposed template matching strategy and the potential ambiguity** is illustrated, where the general steps to realize the best match are labeled (left), while their match confidence measures (right), named Attainable Maximization Likelihood, are plotted to visualize the potential ambiguity for edge cases.

At each iteration, the template-based matching is performed through estimating the image gradient magnitude difference of a 5x5 patch ($S_p = 5$) between the query and the target edge points. The search stops at either maximum growth boundary or discontinuity of edges, where the edge point generating the smallest error is chosen as the best match for local bundle adjustment in Fig. 5 (3).

3) Match confidence based patch size adaption: The major drawback of naive template matching is its ignorance of the potential ambiguity between locally similar edge points. Especially for flat edge cases, like in Fig. 5 (right), where a small pose error may generate large matching bias. Inspired by the research on confidence measures for stereo vision, the best-performance metric in [14], named Attainable Maximum Likelihood (AML), is implemented to distinguish the ambiguous matches along edges as follows:

$$C_m = C_{AML} = \frac{1}{\sum_\lambda \|c_\lambda - c^*\|_2^2}, \tag{11}$$

where c_λ denotes the template matching cost within the search region, and c^* means the cost of the best match.

As long as the C_m is smaller than a pre-defined threshold τ_m, the patch size S_p is increased until the patch size limit τ_S is met. Adaptive patch size allows the algorithm to involve more information for template matching, which improves not only the noise resistance of patch matching [2] but also the accuracy of confidence measure [22].

Noted that we won't perform any match confidence check for points with very small search lengths, where a search length threshold τ_λ is pre-defined. A small search radius implies an accurate pose estimate, so that the best match within this region is also most likely to be the global optima. As a result, we decide to trust the naive matching approach to save computational resources.

4) Depth confidence conditioning: After the patch size adaption process, the matches that still present high ambiguity ($C_m < \tau_m$) will be discarded. Instead of best photometric matches, the edge alignment matches will be passed into the later optimization framework and the depth confidence measure C_d is proposed to predict the their observability at later depth estimation as follows:

$$C_d = \frac{1}{\sigma_\mu} = \frac{cos\theta}{\sigma_{p\|g}} \qquad (12)$$

A small C_d indicates a large potential depth estimation error, which means this particular match is unsuitable for depth estimation, and vice verse. Matches with low match and depth confidences, C_m and C_d, will pass to the local BA layer with a fixed depth at each iteration of joint optimization.

C. Data Association Updates

Inter-frame poses and scene structure may change significantly during local BA optimization so that the matches need to be updated during optimization. To capture the potential erroneous matches, we monitor the reprojection error component $e_{p\perp g}$, calculated from the local BA residuals $\mathbf{r}$. The update conditions can be expressed as follows:

$$e_{p\perp g} = \langle \mathbf{r}, \mathbf{g}_\perp \rangle > k_m \lambda_{max}, \qquad (13)$$

where k_m is the ratio for match update that is typically set to be 0.5 -1.0. It means if the match distance on the edge direction is larger than a certain ratio of maximum search length, the match is most likely to be invalid and needs to be re-associated for estimation accuracy.

V. EVALUATION

In this section, we evaluate our proposed monocular edge VO system quantitatively on publicly available datasets [23] [24] [25] using real-time capable edge detectors [26] [27] [28]. Compared to other edge VO systems concentrating on indoor navigation, our evaluation mainly focuses on challenging outdoor environments, where the sun-glare and pixel over-exposure are the main factors of tracking failure. Our experiments are conducted using a regular laptop featuring an Intel I7 core for our proposed VO pipeline, where the edge detection is calculated on an Nvidia K20 GPU.

A. Evaluation on Edge-Guided Data Association

First, we evaluate the accuracy and efficiency of our proposed edge-guided data association strategy against (1) search length, (2) patch size, and (3) inter-frame distance. The optical flow methods, e.g. the conventional Py-LK [10] and illumination-robust FSDEF [29], serve as the baseline approaches for comparison. Intensity, census, and gradient approaches to template matching are plugged into our framework for evaluation, where the mean translational drift and average processing time are set as metrics for quantitative analysis. A photo-realistic vKITTI [23] dataset with and without simulated illumination changes is used to evaluate our proposed data association approach, where the white Gaussian noise with 1.0 standard variance is added to ground truth inverse depth. We uses 800 points uniformly sampled from edges in the image for evaluations.

The selection of translational drift as the evaluation metric, instead of the more direct average end-point error, is because of the difficulty to generate ground-truth matches for general

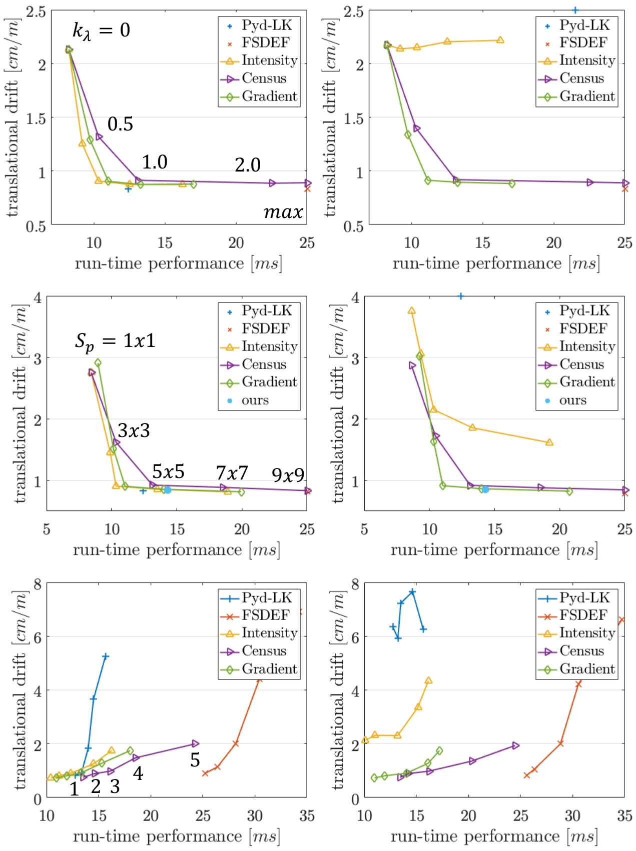

Fig. 6: **Our Proposed Edge-Guided Data Association** is evaluated under regular (left) and illumination-changing (right) sequences. The camera tracking accuracy and run-time performance are evaluated against (1) search length ratio $k_\lambda = \{k_p, k_d\}$ in Eq. (3a) (top), (2) patch size S_p (middle), and (3) inter-frame distances (bottom). It worth noting that the cases that $k_\lambda = 0$ and $k_\lambda = max$ are equivalent to the direct use of edge alignment tracking results and search with a fixed length, respectively.

edge points with sub-pixel precision and the lack of datasets with multi-view dense ground-truth pixel correspondences. Besides, our proposed system is designed to improve the camera tracking accuracy with real-time constraints, which motivates us to carry out *end-to-end* evaluation.

Summarizing observations from Fig. 6: (1) The *gradient* approach shows the best overall performance, in terms of tracking accuracy and efficiency, for regular and illumination-challenging environments. (2) Our proposed search radius formula with k = 1 works well on the tests, which were previously considered as underestimated due to the independence assumptions and multiple approximations. (3) Our proposed adaptive patch size strategy takes less time to realize better correspondences compared with fix-length search strategies. (4) Our proposed edge-guided data association strategy holds the better capability of dealing with large camera motions compared with optical flow methods.

B. Evaluation on Illumination and Motion Robustness

The overall system performance of our proposed edge VO algorithm is evaluated using Symphony Lake [24]

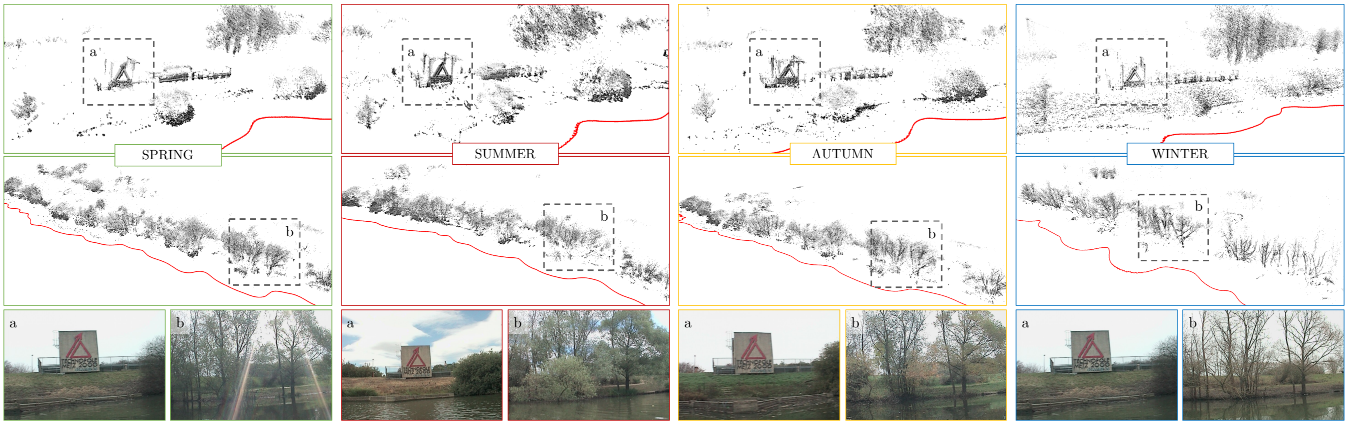

Fig. 7: **Evaluation on Symphony Lake Dataset**. The reconstructed pointclouds and example images of a human-made structure (top) and a natural scene (middle) are presented based on their seasons so that the structural and appearance changes across seasons can be observed. Although we merely present one sun-glare image at SPRING *b*, sun-glares exists in the most test sequences.

TABLE I: System Evaluation against Illumination Changes and Fast Camera Motion.

	Symphony Lake Dataset								Down-Sampled Symphony Lake Dataset								time	
	spring		*summer*		*autumn*		*winter*		*spring*		*summer*		*autumn*		*winter*			
	rate	err	rate	err	rate	err	rate	err	rate	err	rate	err	rate	err	rate	error	track	map
DSO[2]	14.2	17.4	9.3	14.7	2.5	7.7	15.9	21.7	19.9	21.0	12.7	17.8	7.2	11.1	19.4	25.7	58	93
ORB[1]	23.5	26.3	16.4	19.3	8.8	13.3	19.2	24.6	24.1	26.6	16.0	22.1	8.9	13.5	19.6	24.8	-	-
Prev[30]	4.7	9.7	4.4	9.2	**2.2**	6.4	8.0	12.4	10.2	14.7	7.6	11.5	7.3	11.4	14.1	18.5	105	173
Ours(Canny[26])	3.5	8.5	5.0	9.4	2.4	6.6	3.1	7.3	4.2	9.7	6.3	11.5	4.3	9.4	3.5	8.9	63	129
Ours(SE[27])	1.9	7.1	3.2	8.5	2.3	6.5	**2.4**	6.2	2.4	7.4	5.2	10.6	2.4	6.9	2.5	6.5	75	103
Ours(HED[28])	1.9	7.0	3.3	8.4	**2.2**	6.4	2.5	6.4	2.5	7.6	5.2	10.7	2.5	6.7	2.4	6.4	79	106
Hybrid(Canny[26])	3.3	8.1	4.9	9.6	**2.2**	6.4	**2.4**	6.3	3.6	8.9	5.1	10.8	3.4	8.2	2.5	6.4	77	201
Hybrid(SE[27])	**1.8**	**6.8**	3.3	8.4	**2.2**	6.4	**2.4**	**6.1**	**2.1**	**7.0**	4.9	**10.1**	2.3	6.6	**2.3**	**6.2**	89	193
Hybrid(HED[28])	1.9	6.9	**3.1**	**8.2**	**2.2**	**6.3**	**2.4**	**6.1**	2.2	7.1	**4.8**	**10.1**	**2.2**	**6.4**	2.4	**6.2**	92	194

Rates are failure rate per sequence, averaged over 10 trials. Err is a drift rate in $[cm/m]$. Processing time in $[ms]$ per frame.

TABLE II: System Evaluation on Normal Sequences.

KITTI [25]	DSO [2]	ORB [1]	Ours(Canny) [26]	Ours(SE) [27]	Ours(HED) [28]
00	16.83	16.14	18.36	**16.06**	16.11
01	36.32	-	32.59	22.31	**21.55**
02	17.08	**15.58**	17.72	16.77	16.52
03	3.71	**3.44**	4.31	3.64	3.63
04	3.01	3.05	2.93	2.41	**2.33**
05	13.64	**12.96**	13.49	13.05	13.02
06	14.13	13.35	13.44	**12.54**	12.57
07	9.55	9.63	11.36	**8.15**	8.20
08	18.31	**15.43**	19.35	16.24	16.21
09	13.05	12.88	12.73	12.61	**12.47**
Avg.(*)	12.15	11.50	12.63	11.52	**11.44**

$(*)$ excludes KITTI 01 sequence.

dataset, which consists of millions of natural lakeshore images heavily contaminated by (1) smooth (auto-exposure) and (2) sudden (sun-glares) lighting changes, as well as (3) the tree-sky boundary pixel over-exposure. Unlike sun-glares that randomly occurs in sunny days of a year, the over-exposure induced appearance changes show significant variations across seasons. In general, the denser the leaves in Summer, the fewer boundary pixels are 'eaten' by lights the VO system is, therefore, less affected by over-exposure, and the opposite holds in Winter. Based on these observations, we choose 12 surveys (3 surveys per season) heavily contaminated with sun-glare and categorize results based on seasons. For motion robustness evaluation, we down-sample the selected sequences at a sample rate of 3 for fast camera motion simulation. For quantitative analysis, the ground truth pose is calculated through Laser-GPS-based global pose graph optimization described in [31]. The scale of trajectory from monocular VO are corrected using ground truth poses at every 200 frames, while the loop-closure functionality is disabled for ORBSLAM2 for a fair comparison.

In Fig. 7, the point clouds show that our proposed approach reconstructs high-quality human-made structures as well as natural scenes under illumination contamination. Comparing reconstruction results across seasons, we can observe the structural changes of trees between different times of the year. A more detailed quantitative evaluation concerning relative pose errors (RPEs), failure rate, and runtime performance using complete and down-sampled sequences are presented in Table. I. The camera pose error within tracking failure locations is calculated using predefined maximum tracking error (30.0 cm/m). Large errors indicate both low tracking accuracy and low completion rate.

In general, our proposed approach, both with and without hybrid costs, outperforms all other state-of-the-art methods, with the *Winter* sequences showing better performance than the *Summer* ones. Such variation can be attributed to the fact that the *Winter* images generally hold more clear and uniformly distributed edges compared with those of *Summer*. Among different edge detectors, the learned edges (SE [27], HED [28]) present better tracking accuracy and robustness over conventional Canny edge detector, most likely due to its low repeatability for outdoor image sequences [32].

For the evaluation on the influence of high-speed camera motions, both the failure rates and tracking error of state-of-art methods increase significantly after frame subsampling in comparison of our proposed approach. Similar to our previous analysis, the *Winter* sequences using learned edges shows the least performance degradation compared with other approaches. It suggests that well-distributed high-repeatability edges are essential for our proposed edge VO framework, which guarantees both the robustness against illumination changes as well as high-speed camera motions.

Compared with our previous work [30] that incorporated the gradient similarity metric into an optimization framework, this work proposed a probabilistic edge data association strategy to realize a better edge candidate conditioning, resulting in tracking accuracy and robustness boosts without compromising real-time performance. Besides, the hybrid cost optimization further boosts the tracking accuracy for state-of-the-art performance.

C. Evaluation on Regular Sequences

Besides challenging illumination- and motion-challenging dataset, we also evaluate our proposed system on regular sequences for completeness. Table. II shows the absolute trajectory errors (ATEs) after scale correction using ground truth poses on KITTI [25] dataset. Our proposed edge VO approach shows consistent improvements over direct DSO and comparable performance with indirect ORBSLAM2.

VI. CONCLUSIONS

In this work, we propose a monocular edge visual odometry framework, which is a real-time capable algorithm exploiting the edge features and image gradient for illumination-robust camera motion estimation and scene reconstruction. These are obtained by an edge alignment front-end, a finer point correspondence refinement strategy through a fast probabilistic 1D search strategy, and joint optimization in local bundle adjustment. The proposed system successfully overcomes the partial observability issue of monocular edge mapping as well as improving the robustness of outdoor motion estimation. The experimental results indicate that our proposed system outperforms current state-of-art algorithms in terms of illumination- and motion-robustness and shows comparable performance in regular sequences.

REFERENCES

[1] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[2] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[3] L. Kneip, Z. Yi, and H. Li, "Sdicp: Semi-dense tracking based on iterative closest points." in *Proceedings of BMVC*, 2015, pp. 100–1.

[4] Y. Zhou, L. Kneip, and H. Li, "Semi-dense visual odometry for rgb-d cameras using approximate nearest neighbour fields," in *Proceedings of ICRA*. IEEE, 2017, pp. 6261–6268.

[5] Y. Zhou, H. Li, and L. Kneip, "Canny-vo: Visual odometry with rgb-d cameras based on geometric 3-d–2-d edge alignment," *IEEE Transactions on Robotics*, vol. 35, no. 1, pp. 184–199, 2018.

[6] J. Jose Tarrio and S. Pedre, "Realtime edge-based visual odometry for a monocular camera," in *Proceedings of ICCV*, 2015, pp. 702–710.

[7] J. J. Tarrio, C. Smitt, and S. Pedre, "Se-slam: Semi-dense structured edge-based monocular slam," *arXiv preprint arXiv:1909.03917*, 2019.

[8] J.-Y. Bouguet *et al.*, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, no. 1-10, p. 4, 2001.

[9] S. Maity, A. Saha, and B. Bhowmick, "Edge slam: Edge points based monocular visual slam," in *Proceedings of ICCV*, 2017, pp. 2408–2417.

[10] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[11] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual slam," in *Proceedings of ICRA*. IEEE, 2017, pp. 4523–4530.

[12] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 3, p. 24, 2017.

[13] H. Alismail, M. Kaess, B. Browning, and S. Lucey, "Direct visual odometry in low light using binary descriptors," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 444–451, 2016.

[14] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[15] M.-G. Park and K.-J. Yoon, "Leveraging stereo matching with learning-based confidence measures," in *Proceedings of CVPR*, 2015, pp. 101–109.

[16] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *Proceedings of ICCV*. IEEE, 2007, pp. 1–8.

[17] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *Proceedings of ICRA*. IEEE, 2014, pp. 15–22.

[18] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," in *Proceedings of IROS*. IEEE, 2018, pp. 2198–2204.

[19] G. Grisetti, R. Kümmerle, H. Strasdat, and K. Konolige, "g2o: a general framework for (hyper) graph optimization," in *Proceedings of ICRA*, 2011, pp. 9–13.

[20] F. Dellaert and C. Beall, "Gtsam 4.0," *URL: https://bitbucket.org/gtborg/gtsam*, 2017.

[21] J.-L. Blanco, "A tutorial on se (3) transformation parameterizations and on-manifold optimization," *University of Malaga, Tech. Rep*, vol. 3, 2010.

[22] M. Brandao, R. Ferreira, K. Hashimoto, A. Takanishi, and J. Santos-Victor, "On stereo confidence measures for global methods: evaluation, new model and integration into occupancy grids," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 116–128, 2015.

[23] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of CVPR*, 2016, pp. 4340–4349.

[24] S. Griffith, G. Chahine, and C. Pradalier, "Symphony lake dataset," *The International Journal of Robotics Research*, vol. 36, no. 11, pp. 1151–1158, 2017.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of CVPR*, 2012.

[26] J. Canny, "A computational approach to edge detection," in *Readings in computer vision*. Elsevier, 1987.

[27] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proceedings of ICCV*, 2013, pp. 1841–1848.

[28] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of ICCV*, 2015, pp. 1395–1403.

[29] M. Garrigues and A. Manzanera, "Fast semi dense epipolar flow estimation," in *Proceedings of WACV*. IEEE, 2017, pp. 427–435.

[30] X. Wu and C. Pradalier, "Illumination robust monocular direct visual odometry for outdoor environment mapping," in *Proceedings of ICRA*. IEEE, 2019, pp. 2392–2398.

[31] C. Pradalier and F. Pomerleau, "Multi-session lake-shore monitoring in visually challenging conditions," 2018.

[32] X. Wu, A. Benbihi, A. Richard, and C. Pradalier, "Semantic nearest neighbor fields monocular edge visual-odometry," *arXiv preprint arXiv:1904.00738*, 2019.