

AirVO: An Illumination-Robust Point-Line Visual Odometry

Kuan Xu¹, Yuefan Hao², Shenghai Yuan¹, Chen Wang³, Lihua Xie¹, *Fellow, IEEE*

Abstract—This paper proposes an illumination-robust visual odometry (VO) system that incorporates both accelerated learning-based corner point algorithms and an extended line feature algorithm. To be robust to dynamic illumination, the proposed system employs the convolutional neural network (CNN) and graph neural network (GNN) to detect and match reliable and informative corner points. Then point feature matching results and the distribution of point and line features are utilized to match and triangulate lines. By accelerating CNN and GNN parts and optimizing the pipeline, the proposed system is able to run in real-time on low-power embedded platforms. The proposed VO was evaluated on several datasets with varying illumination conditions, and the results show that it outperforms other state-of-the-art VO systems in terms of accuracy and robustness. The open-source nature of the proposed system allows for easy implementation and customization by the research community, enabling further development and improvement of VO for various applications.

I. INTRODUCTION

Due to the good balance in cost and accuracy, VO has been used in an extensive range of applications, especially in the domain of augmented reality and robotics [1]. Despite the existence of numerous well-known works, such as MSCKF [2], VINS-Mono [3] and OKVIS [4], the existing solutions are not robust enough for illumination-challenging conditions [5]. For example, in dynamic illumination environments, visual tracking becomes more challenging and thus the quality of the estimated trajectory is severely affected [6].

On the other hand, deep learning technology has made great progress in many computer vision tasks, which has triggered another research trend [7]. A lot of learning-based feature extraction and matching methods have been proposed and they have been proven to be more robust than handcrafted methods in illumination-challenging environments [8]–[10]. However, they often require huge computational resources and thus are impractical for real-time applications with lightweight robotics platforms such as unmanned aerial vehicles.

Therefore, in this paper, we propose AirVO, an illumination-robust stereo visual odometry. We employ both learning-based feature extraction and matching methods to make our system robust enough in illumination-challenging environments. To

*This research is supported by the National Research Foundation, Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation.

¹Kuan Xu, Shenghai Yuan, and Lihua Xie are with School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, {kuan.xu, shyuan, elhxie}@ntu.edu.sg

²Yuefan Hao is with the Robot R&D Department, Geekplus Corp., Beijing 100107, China yuefan.hao@outlook.com

³Chen Wang is with the Spatial AI & Robotics Lab at The Department of Computer Science & Engineering, State University of New York at Buffalo, Buffalo, NY 14260, USA chenw@sairlab.org

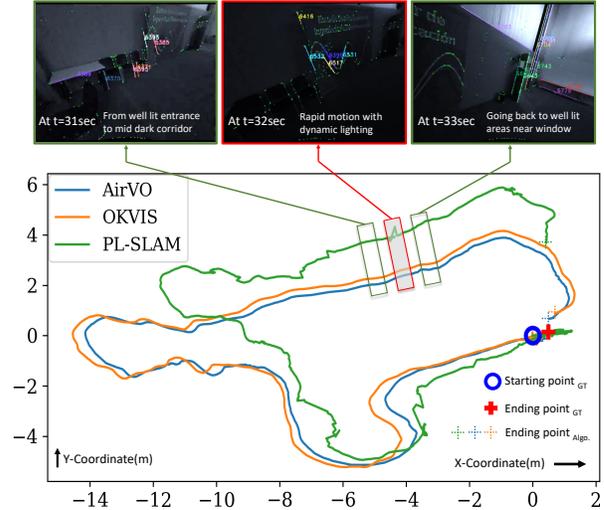


Fig. 1: AirVO is an accurate and robust stereo visual odometry in illumination-challenging environments. More demos are available at https://youtu.be/YfOCL1l_PfU.

achieve real-time and cost-effective performance, we accelerate the CNN and GNN parts and optimize the pipeline, making the feature extraction and tracking five times faster than the original work and the whole system able to run at a rate of about 15Hz on a low-power embedded device.

To improve the accuracy, we also introduce line features into our system. We argue that long lines can provide more stable and accurate constraints, so we merge the short lines detected by LSD [11]. However, line detection is usually unstable in dynamic illumination environments, which makes line tracking and matching more difficult than in good lighting conditions. Besides, line feature triangulation is more difficult than point feature triangulation, because it suffers more from degenerate motion [12]. Therefore, we also propose a fast and efficient illumination-robust line processing pipeline in this paper. Observing that the point tracking in our system is very robust, we associate points with lines according to their distances. Then lines can be matched and triangulated using the matching and triangulation results of related points. The proposed line processing method is shown to be very robust even when the line detection is not stable and the lighting conditions are challenging. It is also very fast due to that it does not need to extract line descriptors. Overall, our contributions are as follows:

- The key contribution in this paper is that we propose a novel hybrid VO system that can effectively handle varying illumination conditions. Our proposed system combines the efficiency of traditional optimization tech-

niques with the robustness of learning-based methods. To our best knowledge, AirVO is the first visual odometry that employs both learning-based feature detection and matching algorithms and can run in real-time on low-power embedded platforms.

- We propose a new line processing pipeline for VO in this paper. Our approach associates 2D lines with learning-based 2D points on the image, leading to more robust feature matching and triangulation. This novel method enhances the accuracy and reliability of VO, especially in illumination-challenging environments.
- We perform extensive experiments that prove the efficiency and effectiveness of the proposed methods. The results show that AirVO outperforms other state-of-the-art VO and visual-inertial odometry (VIO) systems, especially in illumination-challenging environments. Through optimization and acceleration of the relevant parts, our point feature detection and matching achieve more than $5\times$ faster than the original work. Additionally, the system can run at a rate of about 15Hz on a low-power embedded device and 40Hz on a notebook PC. We release source code at <https://github.com/sair-lab/AirVO> to benefit the community.

II. RELATED WORKS

A. Feature Extraction and Tracking for Visual SLAM

Various key-point features have been proposed and applied to different computer vision tasks. Many of these features, e.g., ORB [13], FAST [14], and BRISK [15] are applied to VO and SLAM systems, e.g., ORB-SLAM [16], VINS-Mono [3], because of their balanced effectiveness and efficiency. Two methods are widely used to track the feature points. The first is to use optical flow [3], and the other is matching by descriptor [2], [17]. However, most of the current visual SLAM systems based on the above methods are evaluated in well-lighted environments and make a brightness consistency assumption. Thereby, their performances are significantly affected by challenging lighting conditions, such as dark, over-bright or dynamic illumination conditions.

With the development of deep learning techniques, many learning-based feature extraction and matching methods have been proposed and started to be applied to visual SLAM. Kang *et al.* [18] introduce TFeat network [19] to extract descriptors for FAST corners in a traditional VSLAM pipeline. Tang *et al.* [20] use a neural network to extract robust key-points and binary feature descriptors with the same shape of the ORB. Han *et al.* [21] combine SuperPoint [9] feature extractor with a traditional back-end. Bruno *et al.* proposed LIFT-SLAM [22], where they use LIFT [8] to extract features. Li *et al.* [23] replace ORB feature with SuperPoint in ORB-SLAM2 and optimize the feature extraction with Intel OpenVINO toolkit. However, the above methods still adopt traditional methods to track or match these learning-based features, making them not robust enough to changing illumination. Sarlin *et al.* propose HF-Net [24], where they integrate SuperPoint and SuperGlue [10] into COLMAP [25], a structure from motion

software. HF-Net achieves good performance for visual place recognition tasks but requires huge computing resources and is unable to build maps in real-time.

Unlike current methods, AirVO introduces both learning-based feature extraction and matching methods in the VO system, which makes our system robust enough in illumination-challenging environments. By accelerating the CNN and GNN parts, our system can perform pose estimation and build maps in real-time on low-power platforms.

B. Line Matching for Visual SLAM

Line features widely exist in man-made environments, which can provide additional constraints. One of the challenges of using lines in visual SLAM is to perform line matching. A method used in many current SLAM systems [12], [26]–[28] is to match lines through LBD [29] descriptor. This method may make the line matching fail as the traditional line detection method, such as LSD [11], may be unstable. To overcome this, some systems [30]–[32] sample some points on a line, and then track the line by tracking these points. However, using either minimizing photometric error along the epipolar line or zero-normalized cross-correlation (ZNCC) matching method [33] cannot ensure robust line tracking in dynamic illumination environments.

C. Visual SLAM for Dynamic Illumination

Several methods have been proposed to improve the robustness of VO and Visual SLAM to illumination changes. DSO [34] models brightness changes and jointly optimizes camera poses and photometric parameters. DRMS [35] and AFE-ORB-SLAM [36] utilize various image enhancements. Some systems try different methods, such as ZNCC, locally-scaled sum of squared differences (LSSD) and dense descriptor computation, to achieve robust tracking [37]–[39]. These methods mainly focus on either global or local illumination change for all kinds of images, however, lighting conditions often affect the scene differently in different areas [40].

Other related methods include that of Huang and Liu [41], which presents a multi-feature extraction algorithm to extract two kinds of image features when a single-feature algorithm fails to extract enough feature points. Kim *et al.* [42] employ a patch-based affine illumination model during direct motion estimation. Chen *et al.* [43] minimize the normalized information distance (NID) with nonlinear least square optimization for image registration. Alismail *et al.* propose a binary feature descriptor using a descriptor assumption to avoid the brightness constancy [44].

III. METHODOLOGY

A. System Overview

The proposed framework is shown in Fig. 2. It is a hybrid VO system where we utilize both the learning-based front end and the traditional optimization backend. For each stereo image pair, we first employ SuperPoint [9] to extract feature points on the left image and match them with the last keyframe using SuperGlue [10], and in parallel, we also extract line features. Then the two kinds of features

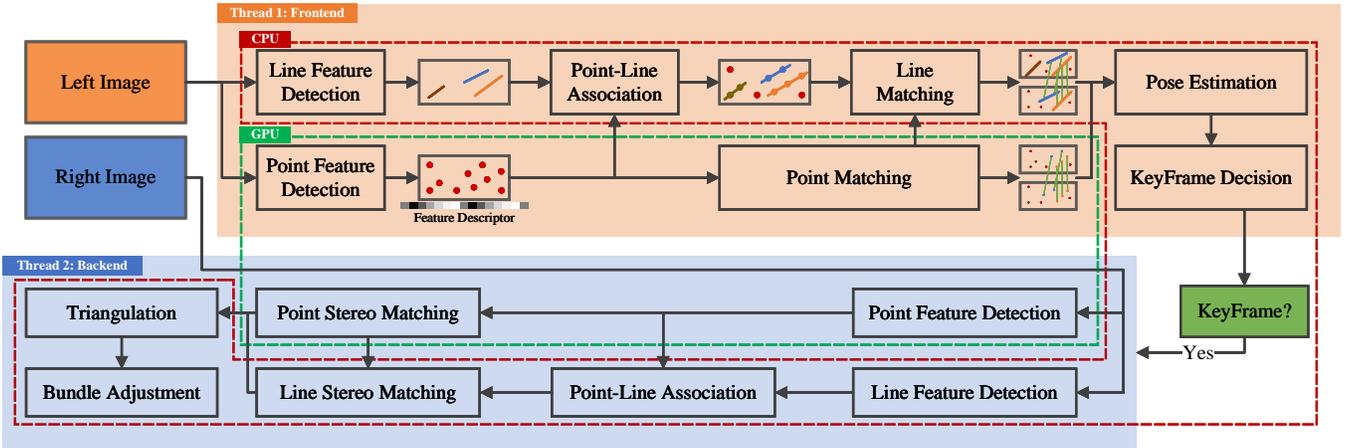


Fig. 2: The framework of AirVO. The system is split into two main threads, which are represented by two different colored regions. The modules in the red dotted box and green dotted box run on the CPU and GPU, respectively.

are associated according to their distances and line features are matched using the matching results of associated points. After that, we perform an initial pose estimation and reject outliers. Based on the results, we select keyframes, extract features on the right image and triangulate 2D points and lines of keyframes. Finally, the local bundle adjustment will be performed to optimize points, lines and keyframe poses.

To improve system efficiency, We replace the 32-bit floating-point arithmetic of CNNs and GNNs in our system with 16-bit floating-point arithmetic, which makes feature extraction and tracking more than five times faster than the original code on the embedded device. We also design a multi-thread pipeline that utilizes both CPU and GPU resources. A producer-consumer model is used to split the system into two main threads, i.e., the feature thread and the optimization thread. In the feature thread, we use two sub-threads to process point features and line features separately. In one sub-thread, the point feature extraction and matching with the last frame are put on the GPU while in parallel, the other sub-thread is used to extract line features on the CPU. In the optimization thread, we perform initial pose estimation and keyframe decision. If a new keyframe is selected, we extract both point and line features on its right image and optimize its pose with a local map.

B. 2D Line Processing

We first give the details of 2D line processing in our system, which includes line segment detection and matching.

1) *Detection*: Line detection of AirVO is based on a traditional method (i.e., LSD [11]) for efficiency. LSD is a popular line detection algorithm. However, it suffers from the problem of dividing a line into multiple segments. Therefore, we improve it by merging two line segments \mathbf{l}_1 and \mathbf{l}_2 if the following conditions are satisfied:

- The angle between \mathbf{l}_1 and \mathbf{l}_2 is less than a threshold δ_θ .
- The distance between the midpoint of one line and the other line is not greater than a certain value δ_d .
- If projections of \mathbf{l}_1 and \mathbf{l}_2 on \mathbf{X} -coordinate axis and \mathbf{Y} -coordinate axis do not have overlap, the distance of the two closest endpoints is smaller than a threshold δ_{ep} .



Fig. 3: Lines detected by LSD (left) and by AirVO (right). We merge unstable short lines into stable longer lines.

The line features detected in our system and comparison with LSD are shown in Fig. 3. We argue that long line segments are more repetitive and less affected by noise than the short ones, so after the merger, line segments whose lengths are less than a preset threshold will be filtered out so that only long line segments are used in the following stages.

2) *Matching*: Most of the current VO and SLAM systems use LBD algorithm or tracking sample points to match or track lines. LBD algorithm extracts the descriptor from a local band region of the line, so it suffers from unstable line detection in dynamic illumination environments where the line length may change and thus the local band region would be different between two frames. Tracking sample points can track the line which has different lengths in two frames, but current SLAM systems usually use optical flow to track the sample points, which have a bad performance when the light conditions change rapidly or violently. Some learning-based line feature matching methods [45], [46] are also proposed, however, they are rarely used in current SLAM systems as a result of the requirement for huge computational resources. We do not employ them either because it is difficult to make the system run in real-time on low-power embedded platforms if both learning-based point features and learning-based line features are used simultaneously.

Therefore, to address both the effectiveness problem and efficiency problem, we design a fast and robust line-matching method for dynamic illumination environments. First, we associate point features with line segments through the distances between points and lines. Assume that M key-points and N line segments are detected on the image, where each point is denoted as $\mathbf{p}_i = (x_i, y_i)$ and each line segment is denoted as

$\mathbf{l}_j = (A_j, B_j, C_j, x_{j,1}, y_{j,1}, x_{j,2}, y_{j,2})$, where (A_j, B_j, C_j) are line parameters of \mathbf{l}_j and $(x_{j,1}, y_{j,1}, x_{j,2}, y_{j,2})$ are the endpoints. We first compute the distance between \mathbf{p}_i and \mathbf{l}_j through:

$$d_{ij} = d(\mathbf{p}_i, \mathbf{l}_j) = \frac{|A_j \cdot x_i + B_j \cdot y_i + C_j|}{\sqrt{A_j^2 + B_j^2}}. \quad (1)$$

If $d_{ij} < 3$ and the projection of \mathbf{p}_i on the coordinate axis lies within the projections of line segment endpoints, i.e., $\min(x_{j,1}, x_{j,2}) \leq x_i \leq \max(x_{j,1}, x_{j,2})$ or $\min(y_{j,1}, y_{j,2}) \leq y_i \leq \max(y_{j,1}, y_{j,2})$, we will say \mathbf{p}_i belongs to \mathbf{l}_j . Then the line segments on two images can be matched based on the point-matching result of these two images. For ${}^k\mathbf{l}_m$ on image k and ${}^{k+1}\mathbf{l}_n$ on image $k+1$, we compute a score to represent the confidence of that they are the same line:

$$S_{mn} = \frac{N_{pm}}{\min({}^kN_m, {}^{k+1}N_n)}, \quad (2)$$

where N_{pm} is the matching number between point features belonging to ${}^k\mathbf{l}_m$ and point features belonging to ${}^{k+1}\mathbf{l}_n$. kN_m and ${}^{k+1}N_n$ are the numbers of point features belonging to ${}^k\mathbf{l}_m$ and ${}^{k+1}\mathbf{l}_n$, respectively. Then if $S_{mn} > \delta_S$ and $N_{pm} > \delta_N$, where δ_S and δ_N are two preset thresholds, we will regard ${}^k\mathbf{l}_m$ and ${}^{k+1}\mathbf{l}_n$ as the same line.

Because the point matching is illumination-robust and feature association is not affected by lighting changes, the proposed line tracking method is very robust to dynamic illumination environments, as shown in Fig. 4.

C. 3D Line Processing

In this part, we will introduce our 3D line processing methods. Compared with 3D points, 3D lines have more degrees of freedom, so we first introduce their representations in different stages. Then the methods of line triangulation, i.e., constructing a 3D line from some 2D line segments, and line re-projection, i.e., projecting the 3D line to the image plane, will be illustrated in detail.

1) *Representation*: We use Plücker coordinates [47] to represent a 3D spatial line:

$$\mathbf{L} = \begin{bmatrix} \mathbf{n} \\ \mathbf{v} \end{bmatrix} \in \mathbb{R}^6, \quad (3)$$

where \mathbf{v} is the direction vector of the line and \mathbf{n} is the normal vector of the plane determined by the line and the origin. Plücker coordinates are used for 3D line triangulation, transformation, and projection to the image. It is over-parameterized because it is a 6-dimensional vector, but a 3D line has only four degrees of freedom. In the graph optimization stage, the extra degrees of freedom will increase the computational cost and cause the numerical instability of the system [27]. Therefore, we also use orthonormal representation [47] to represent a 3D line:

$$(\mathbf{U}, \mathbf{W}) \in SO(3) \times SO(2) \quad (4)$$

The relationship between Plücker coordinates and orthonormal representation is similar to $SO(3)$ and $so(3)$. Orthonormal



Fig. 4: Line matching of AirVO in challenging scenes. Matched lines are drawn in the same color. Circles on a line are the points associated with the line. A larger radius indicates that the point is associated with more lines.

representation can be obtained from Plücker coordinates by:

$$\mathbf{L} = [\mathbf{n} \mid \mathbf{v}] = \underbrace{\begin{bmatrix} \frac{\mathbf{n}}{\|\mathbf{n}\|} & \frac{\mathbf{v}}{\|\mathbf{v}\|} & \frac{\mathbf{n} \times \mathbf{v}}{\|\mathbf{n} \times \mathbf{v}\|} \end{bmatrix}}_{\mathbf{U} \in SO(3)} \underbrace{\begin{bmatrix} \|\mathbf{n}\| & & \\ & \|\mathbf{v}\| & \\ & & \end{bmatrix}}_{\Sigma_{3 \times 2}}, \quad (5)$$

where $\Sigma_{3 \times 2}$ is a diagonal matrix and its two non-zero entries defined up to scale can be represented by an $SO(2)$ matrix, i.e., \mathbf{W} . In practice, this conversion can be done simply and quickly with the QR decomposition.

2) *Triangulation*: Triangulation is to initialize a 3D line from two or more 2D line observations. Two methods are used to triangulate a 3D line in our system. The first is similar to the line triangulation algorithm B in [12], where a 3D line can be computed from two planes. To achieve this, we select two line segments, \mathbf{l}_1 and \mathbf{l}_2 , on two images, which are two observations of a 3D line. \mathbf{l}_1 and \mathbf{l}_2 can be back-projected and construct two 3D planes, π_1 and π_2 . Then the 3D line can be regarded as the intersection of π_1 and π_2 .

However, triangulating a 3D line is more difficult than triangulating a 3D point, because it suffers more from degenerate motions [12]. Therefore, we also employ a second line triangulation method if the above method fails, where points are utilized to compute the 3D line. In Section III-B.2, we have associated point features with line features. So to initialize a 3D line, two triangulated points \mathbf{X}_1 and \mathbf{X}_2 , which belong to this line and have the shortest distance from this line on the image plane are selected. Then the Plücker coordinates of this line can be obtained through:

$$\mathbf{L} = \begin{bmatrix} \mathbf{n} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \times \mathbf{X}_2 \\ \frac{\mathbf{X}_1 - \mathbf{X}_2}{\|\mathbf{X}_1 - \mathbf{X}_2\|} \end{bmatrix}. \quad (6)$$

Because the selected 3D points have been triangulated in the point triangulating stage, this method requires little extra computation. It is very efficient and robust.

3) *Re-projection*: We use Plücker coordinates to transform and re-project 3D lines. First, we convert the 3D line from

TABLE I: Translational error (RMSE) without loop closing and re-localization on the OIVIO dataset (unit: m). **L** refers to tracking lost and **D** refers to sequences where RMSEs are larger than 10m.

| Sequence | VINS-Fusion | StructVIO | UV-SLAM | PL-SLAM | OKVIS | ORB-SLAM2 | Basalt-VIO | AirVO |
|--------------|-------------|-----------|----------|---------|---------------|---------------|------------|---------------|
| MN_015_GV_01 | 0.1033 | 8.6098 | 0.4991 | 1.3166 | <u>0.0663</u> | 0.0762 | 0.2157 | 0.0537 |
| MN_015_GV_02 | D | L | D | 0.9523 | 1.5320 | <u>0.0776</u> | 0.1533 | 0.0619 |
| MN_050_GV_01 | D | D | D | 1.1538 | 0.7785 | <u>0.0839</u> | 0.1857 | 0.0756 |
| MN_050_GV_02 | D | D | D | 1.0055 | 0.7385 | <u>0.0755</u> | 0.1026 | 0.0717 |
| MN_100_GV_01 | D | D | D | 0.8455 | 0.8729 | <u>0.0892</u> | 0.1965 | 0.0646 |
| MN_100_GV_02 | D | D | D | 0.6708 | 0.4360 | <u>0.0848</u> | 0.0922 | 0.0770 |
| TN_015_GV_01 | 0.1541 | 7.5849 | 1.6695 | 1.8856 | 0.3063 | 0.0902 | 0.1478 | <u>0.1009</u> |
| TN_050_GV_01 | 0.2079 | D | 2.5948 | 1.9335 | 0.2262 | 0.0965 | 0.5214 | <u>0.0971</u> |
| TN_100_GV_01 | 0.4063 | D | 1.4496 | 1.5263 | 0.3984 | <u>0.1044</u> | 0.1162 | 0.0578 |

the world frame to the camera frame:

$${}^c\mathbf{L} = \begin{bmatrix} {}^c\mathbf{n} \\ {}^c\mathbf{v} \end{bmatrix} = \begin{bmatrix} {}^c_w\mathbf{R} & [{}^c_w\mathbf{t}]_{\times} {}^c_w\mathbf{R} \\ \mathbf{0} & {}^c_w\mathbf{R} \end{bmatrix} \begin{bmatrix} {}^w\mathbf{n} \\ {}^w\mathbf{v} \end{bmatrix} = {}^c_w\mathbf{H} {}^w\mathbf{L}, \quad (7)$$

where ${}^c\mathbf{L}$ and ${}^w\mathbf{L}$ are Plücker coordinates of 3D line in the camera frame and world frame, respectively. ${}^c_w\mathbf{R} \in SO(3)$ is the rotation matrix from world frame to camera frame and ${}^c_w\mathbf{t} \in \mathbb{R}^3$ is the translation vector. $[\cdot]_{\times}$ denotes the skew-symmetric matrix of a vector and ${}^c_w\mathbf{H}$ is the transformation matrix of 3D lines from world frame to camera frame.

Then the 3D line ${}^c\mathbf{L}$ can be projected to the image plane through a line projection matrix ${}^i_c\mathbf{P}$:

$${}^i\mathbf{l} = \begin{bmatrix} A \\ B \\ C \end{bmatrix} = {}^i_c\mathbf{P} {}^c\mathbf{L}_{[:3]} = \begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ -f_y c_x & -f_x c_y & f_x f_y \end{bmatrix} {}^c\mathbf{n}, \quad (8)$$

where ${}^i\mathbf{l} = [A \ B \ C]^T$ is the re-projected 2D line on image plane. ${}^c\mathbf{L}_{[:3]}$ denotes the first three rows of vector ${}^c\mathbf{L}$.

D. Keyframe Selection

Observing that the learning-based data association method used in our system is able to track two frames that have a large baseline, so different from the frame-by-frame tracking strategy used in other VO or visual SLAM systems, we only match the current frame with the last keyframe, as this can reduce the tracking error. A frame will be selected as a keyframe if any of the following conditions is satisfied:

- The distance to the last keyframe is larger than δ_d^{kf} .
- The angle with the last keyframe is larger than δ_θ^{kf} .
- The number of tracked map-points is smaller than N_1^{kf} and bigger than N_2^{kf} , where $N_2^{kf} < N_1^{kf}$.
- Tracked map-points are more than N_2^{kf} but the tracking-lost happened in the last frame, i.e., map-points tracked by the last frame are less than N_2^{kf} .

where δ_d^{kf} , δ_θ^{kf} , N_1^{kf} and N_2^{kf} are all preset thresholds.

E. Graph Optimization

We select N_{kf}^{go} keyframes and construct a co-visibility graph similar to ORB-SLAM [16], where map points, 3D lines, and keyframes are vertices and constraints are edges. Both point constraints and line constraints are used in our system and the related error terms are defined as follows.

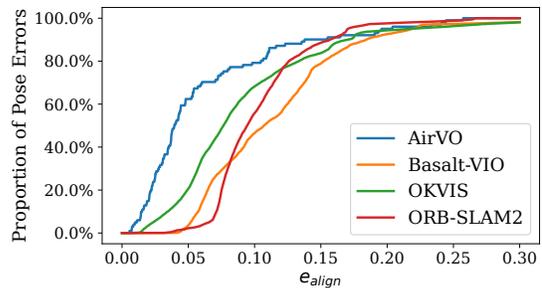


Fig. 5: Comparison based on the OIVIO dataset. The vertical axis is the proportion of pose errors that are less than the given alignment error threshold on the horizontal axis.

1) *Line Re-projection Error*: If the frame k can observe the 3D line ${}^w\mathbf{L}_i$, then the re-projection error is defined as:

$$\mathbf{E}_{l_{k,i}} = e_l \left({}^k\bar{\mathbf{l}}_i, {}^k\mathbf{P} ({}^c_w\mathbf{H} {}^w\mathbf{L}_i)_{[:3]} \right) \in \mathbb{R}^2, \quad (9a)$$

$$e_l \left({}^k\bar{\mathbf{l}}_i, {}^k\mathbf{l}_i \right) = \left[d \left({}^k\bar{\mathbf{p}}_{i,1}, {}^k\mathbf{l}_i \right) \ d \left({}^k\bar{\mathbf{p}}_{i,2}, {}^k\mathbf{l}_i \right) \right]^T, \quad (9b)$$

where ${}^k\bar{\mathbf{l}}_i$ is the observation of ${}^w\mathbf{L}_i$ on frame k , $d(\mathbf{p}, \mathbf{l})$ is the distance between point \mathbf{p} and line \mathbf{l} , and ${}^k\bar{\mathbf{p}}_{i,1}$ and ${}^k\bar{\mathbf{p}}_{i,2}$ are the endpoints of ${}^k\bar{\mathbf{l}}_i$.

2) *Point Re-projection Error*: If the frame k can observe the 3D point ${}^w\mathbf{X}_q$, then the re-projection error is defined as:

$$\mathbf{E}_{p_{k,q}} = {}^k\bar{\mathbf{x}}_q - \pi \left({}^c_w\mathbf{R} {}^w\mathbf{X}_q + {}^c_w\mathbf{t} \right), \quad (10)$$

where ${}^k\bar{\mathbf{x}}_q$ is the observation of ${}^w\mathbf{X}_q$ on frame k and $\pi(\cdot)$ represents the camera projection.

IV. EXPERIMENTS

In this section, experimental results will be presented to demonstrate the performance of our method. We take pre-trained SuperPoint and SuperGlue to detect and match feature points without any fine-tuning training. The experiments are conducted on two datasets: OIVIO dataset [48] and UMA visual-inertial dataset [6]. To prove the efficiency of the proposed line processing pipeline, we compare AirVO with state-of-the-art point-line VO and visual SLAM systems, i.e., PL-SLAM [26], StructVIO [31] and UV-SLAM [28]. PL-SLAM and UV-SLAM use the LBD descriptor to match line features while StructVIO tracks line features by tracking sampling points on lines. We also add stereo-mode VINS-Fusion [3], ORB-SLAM2 [17], Basalt-VIO [39] and VIO-mode OKVIS [49] to the baselines. To handle the dynamic illumination problem, VINS-Fusion uses a failure detection

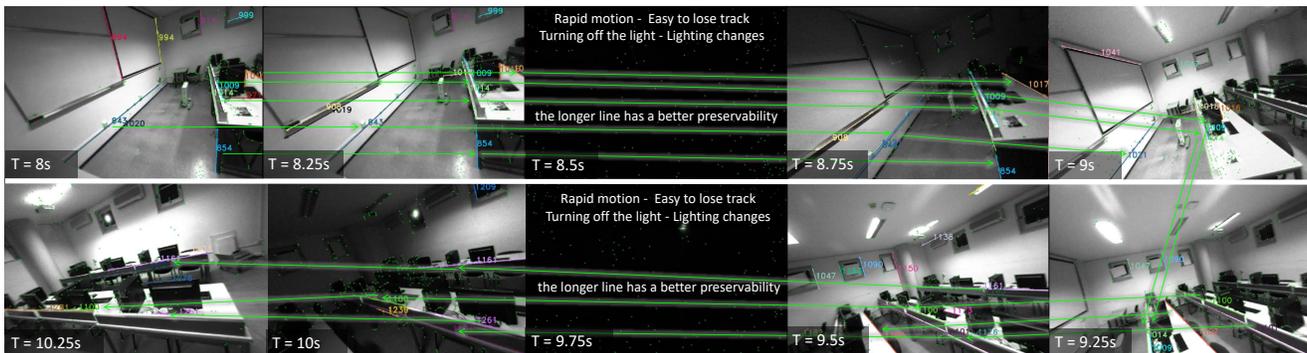


Fig. 6: A challenging sequence in UMA-VI dataset with significant illumination changes. The image may suddenly go dark as a result of turning off the lights, which is very difficult for feature tracking.

TABLE II: Translational error (RMSE) on the UMA-VI dataset (unit: m). The best results are **highlighted**.

| Sequence | PL-SLAM | OKVIS | AirVO |
|------------------|---------|--------|---------------|
| conference-csc1 | 2.6974 | 1.1181 | 0.5236 |
| conference-csc2 | 1.5956 | 0.4696 | 0.1607 |
| third-floor-csc1 | 4.4779 | 0.2525 | 0.1760 |
| third-floor-csc2 | 6.0675 | 0.2161 | 0.1312 |
| average | 3.7096 | 0.5141 | 0.2479 |

and recovery module, while StructVIO uses the ZNCC method and Basalt-VIO uses the LSSD KLT method. The following experiments will prove that AirVO outperforms these methods in illumination-challenging environments. As the proposed method is a VO system, we disabled the loop closure part and re-localization from the above baselines.

Note that as a result of lacking the ground truth of line matching and triangulation in dynamic illumination environments, we prove the effectiveness of the proposed line processing method by comparing it with other point-line systems and an ablation study instead of designing an extra line matching or triangulation comparison. Like [50]–[52], we compare with ORB-SLAM2 instead of ORB-SLAM3 [53] because the newly added atlas and IMU in ORB-SLAM3 are unfair to compare with the visual-only odometry, and they are difficult to remove because of the high coupling system. We also do not add DX-SLAM [23] and GCNv2-SLAM [20] to the baselines since they are based on RGB-D inputs and thus cannot run on the stereo datasets.

A. Results on OIVIO Benchmark

OIVIO dataset collects visual-inertial data in tunnels and mines. In each sequence, the scene is illuminated by an onboard light of approximately 1300, 4500, or 9000 lumens. We used all nine sequences with ground truth acquired by the Leica TCRP1203 R300. The performance of translational error is presented in Table I. The two most accurate results are **highlighted** and underlined, respectively. AirVO achieves the best performance on 7 sequences and the second-best performance on the other 2 sequences, which outperforms other state-of-the-art algorithms. We notice that VINS-Fusion, StructVIO and UV-SLAM lost track on many sequences, this may be because their feature tracking methods, i.e., LSSD KLT sparse optical flow, ZNCC, LBD descriptor, are not

TABLE III: Ablation study. Translational error (RMSE) of AirVO^{w/o line} and AirVO on the UMA-VI and OIVIO datasets (unit: m). The best results are **highlighted**.

| | Sequence | AirVO ^{w/o line} | AirVO |
|--------------|------------------|---------------------------|---------------|
| UMA-VI | conference-csc1 | 2.4789 | 0.5236 |
| | conference-csc2 | 0.2323 | 0.1607 |
| | third-floor-csc1 | 0.1736 | 0.1760 |
| | third-floor-csc2 | 0.1629 | 0.1312 |
| OIVIO | MN_015_GV_01 | 0.1035 | 0.0537 |
| | MN_015_GV_02 | 0.0668 | 0.0619 |
| | MN_050_GV_01 | 0.1051 | 0.0756 |
| | MN_050_GV_02 | 0.1049 | 0.0717 |
| | MN_100_GV_01 | 0.1177 | 0.0646 |
| | MN_100_GV_02 | 0.0921 | 0.0770 |
| | TN_015_GV_01 | 0.1155 | 0.1009 |
| | TN_050_GV_01 | 0.0987 | 0.0971 |
| TN_100_GV_01 | 0.0879 | 0.0578 | |

robust enough in illumination-challenging environments.

We show a comparison of our method with selected baselines on OIVIO TN_100_GV_01 sequence in Fig. 5. In this case, the robot goes through a mine with onboard illumination. The distance is about 150 meters and the average speed is about 0.84m/s. The plot shows the proportion of pose errors on the horizontal axis that are less than the given alignment error threshold on the horizontal axis. AirVO achieves the most accurate result on this sequence.

B. Results on UMA-VI Benchmark

UMA-VI dataset is a visual-inertial dataset gathered in illumination-challenging scenarios with handheld custom sensors. We selected sequences with illumination changes to evaluate our system. As shown in Fig. 6, it contains many sub-sequences where the image suddenly goes dark as a result of turning off the lights. It is more challenging than OIVIO dataset, so we only select methods proved to be illumination-robust in Section IV-A as baselines, i.e., ORB-SLAM2, PL-SLAM, OKVIS and Basalt-VIO. The translational errors are presented in Table II. As ORB-SLAM2 and Basalt-VIO lost track on all 4 sequences, we do not list their results. It can be seen that AirVO outperforms other methods. Its average translational error is only 6.7% of PL-SLAM and 48.2% of OKVIS. We notice that the aligned errors are larger than those on the OIVIO dataset. It is because the UMA-VI dataset only gives the ground truth of the beginning and end

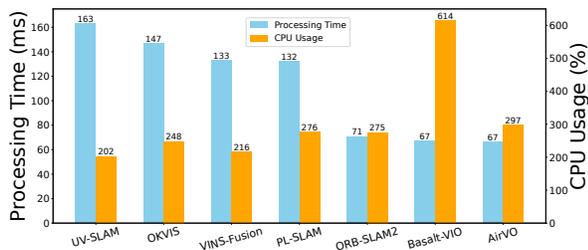


Fig. 7: Bar chart showing the efficiency of different algorithms, as measured by CPU usage (%) and per-frame processing time (ms) on Nvidia Jetson AGX Xavier (2018).

TABLE IV: The average running time comparison of principal components with PL-SLAM.

| | PE | LE | PM | LM | IPE | BA |
|---------|-------|-------|-------|-------|------|--------|
| PL-SLAM | 70 ms | 96 ms | 1 ms | 29 ms | 1 ms | 101 ms |
| AirVO | 25 ms | 29 ms | 10 ms | 2 ms | 4 ms | 264 ms |

of each sequence, which makes the errors appear larger, and the scenes are more difficult for VO or VIO systems.

We also compare the trajectory of AirVO with OKVIS and PL-SLAM on conference-csc2 sequence as shown in Fig. 1. The traveling distance of this sequence is about 50 meters and the average speed is about 0.75m/s. It clearly shows that AirVO produces the best accuracy in this challenging case. The drift error of AirVO is about 1.0%. OKVIS and PL-SLAM are 1.5% and 7.1%, respectively.

C. Ablation Study

To show the effectiveness of the proposed line processing method, we remove line features from AirVO and name it as AirVO^{w/o line}. The comparison results of AirVO and AirVO^{w/o line} on OIVIO and UMA-VI datasets are presented in Table III. It can be seen that AirVO outperforms AirVO^{w/o line} on 12 of 13 sequences, and utilizing line features reduces the translational error by 58.2% on average, which demonstrates that the proposed line processing method can improve the performance of the system.

D. Runtime Analysis

This section presents the running time analysis of the proposed system. The evaluation is performed on the Nvidia Jetson AGX Xavier (2018), which is a low-power embedded platform with an 8-core ARM v8.2 64-bit CPU and a low-power 512-core NVIDIA Volta GPU. The resolution of the input image sequence is 640 × 480. For all algorithms, we extract 200 points and disabled the loop closure, re-localization and visualization part for a fair comparison.

1) *CNN and GNN Acceleration*: We first verify the acceleration of the point detection and matching network. In our system, detecting and tracking feature points for one image take 64 ms, while the original code needs 342 ms. So it is about 5.3× faster than the origin.

2) *Efficiency Comparison*: We also compare the algorithm efficiency, as measured by CPU usage and per-frame processing time. The result is presented in Fig. 7. It can be seen that AirVO is one of the fastest methods (about 15 FPS) while

the CPU usage is roughly the same as other methods because of utilizing the GPU resources. Notice that only the binary executable file of Struct-VIO is available, which is compiled on the x86 computer and cannot run on the Jetson platform, so we did not add it to this comparison.

3) *Detailed Running Time*: We also present the detailed running time of each module of PL-SLAM and AirVO in Table IV, where PE is point extraction, LE is line extraction, PM is point matching, LM is line matching, IPE is initial pose estimation and BA is keyframe processing and local bundle adjustment. It can be seen that the line processing pipeline of AirVO is much more efficient than PL-SLAM. Our BA module has higher runtime than PL-SLAM, this may be because more stable features are detected, tracked and taken into optimization in our system. Notice that these modules run in parallel and BA module is a non-real-time back-end thread, so the running time of the whole system is not the simple accumulation of each module.

V. CONCLUSIONS

In this work, we presented an illumination-robust visual odometry based on learning-based key-point detection and matching methods. To improve the accuracy, line features are also utilized in our system. We proposed a novel line processing pipeline to make line tracking robust enough in illumination-dynamic environments. In the experiments, we showed that the proposed method achieved superior performance in illumination-dynamic environments and could run in real-time on low-power devices. We open the source code to benefit the robotic community. For future work, we will extend AirVO to a SLAM system by adding loop closing, re-localization and map reuse. We hope to build an illumination-robust visual map for long-term localization.

REFERENCES

- [1] A. Macario Barros, M. Michel, Y. Moline, G. Corre, and F. Carrel, “A comprehensive survey of visual slam algorithms,” *Robotics*, vol. 11, no. 1, p. 24, 2022.
- [2] A. I. Mourikis and S. I. Roumeliotis, “A multi-state constraint kalman filter for vision-aided inertial navigation,” in *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE, 2007, pp. 3565–3572.
- [3] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.
- [4] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, “Keyframe-based visual-inertial odometry using nonlinear optimization,” *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.
- [5] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.
- [6] D. Zuñiga-Noël, A. Jaenal, R. Gomez-Ojeda, and J. Gonzalez-Jimenez, “The uma-vi dataset: Visual-inertial odometry in low-textured and dynamic illumination environments,” *The International Journal of Robotics Research*, vol. 39, no. 9, pp. 1052–1060, 2020.
- [7] K. Xu, C. Wang, C. Chen, W. Wu, and S. Scherer, “Aircode: A robust object encoding method,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1816–1823, 2022.
- [8] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “Lift: Learned invariant feature transform,” in *European conference on computer vision*. Springer, 2016, pp. 467–483.

- [9] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
- [10] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.
- [11] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A line segment detector," *Image Processing On Line*, vol. 2, pp. 35–55, 2012.
- [12] Y. Yang, P. Geneva, K. Eickenhoff, and G. Huang, "Visual-inertial odometry with point and line features," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2447–2454.
- [13] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2564–2571.
- [14] D. G. Viswanathan, "Features from accelerated segment test (fast)," in *Proceedings of the 10th workshop on image analysis for multimedia interactive services, London, UK*, 2009, pp. 6–8.
- [15] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2548–2555.
- [16] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.
- [17] R. Mur-Artal and J. D. Tardos, "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.
- [18] R. Kang, J. Shi, X. Li, Y. Liu, and X. Liu, "Df-slam: A deep-learning enhanced visual slam system based on deep local features," *arXiv preprint arXiv:1901.07223*, 2019.
- [19] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Bmvc*, vol. 1, no. 2, 2016, p. 3.
- [20] J. Tang, L. Ericsson, J. Folkesson, and P. Jensfelt, "Gcnv2: Efficient correspondence prediction for real-time slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3505–3512, 2019.
- [21] X. Han, Y. Tao, Z. Li, R. Cen, and F. Xue, "Superpointvo: A lightweight visual odometry based on cnn feature extraction," in *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE)*. IEEE, 2020, pp. 685–691.
- [22] H. M. S. Bruno and E. L. Colombini, "Lift-slam: A deep-learning feature-based monocular visual slam method," *Neurocomputing*, vol. 455, pp. 97–110, 2021.
- [23] D. Li, X. Shi, Q. Long, S. Liu, W. Yang, F. Wang, Q. Wei, and F. Qiao, "Dxslam: A robust and efficient visual slam system with deep features," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4958–4965.
- [24] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12716–12725.
- [25] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.
- [26] R. Gomez-Ojeda, F.-A. Moreno, D. Zuniga-Noël, D. Scaramuzza, and J. Gonzalez-Jimenez, "Pl-slam: A stereo slam system through the combination of points and line segments," *IEEE Transactions on Robotics*, vol. 35, no. 3, pp. 734–746, 2019.
- [27] X. Zuo, X. Xie, Y. Liu, and G. Huang, "Robust visual slam with point and line features," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1775–1782.
- [28] H. Lim, J. Jeon, and H. Myung, "UV-SLAM: Unconstrained line-based slam using vanishing points for structural mapping," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1518–1525, 2022.
- [29] L. Zhang and R. Koch, "An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794–805, 2013.
- [30] L. Zhou, S. Wang, and M. Kaess, "Dplvo: Direct point-line monocular visual odometry," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7113–7120, 2021.
- [31] D. Zou, Y. Wu, L. Pei, H. Ling, and W. Yu, "Structvio: visual-inertial odometry with structural regularity of man-made environments," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 999–1013, 2019.
- [32] B. Xu, P. Wang, Y. He, Y. Chen, Y. Chen, and M. Zhou, "Leveraging structural information to improve point line visual-inertial odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3483–3490, 2022.
- [33] L. Di Stefano, S. Mattocchia, and F. Tombari, "Zncc-based template matching using bounded partial correlation," *Pattern recognition letters*, vol. 26, no. 14, pp. 2129–2134, 2005.
- [34] 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.
- [35] Q. Gu, P. Liu, J. Zhou, X. Peng, and Y. Zhang, "Drms: Dim-light robust monocular simultaneous localization and mapping," in *2021 International Conference on Computer, Control and Robotics (ICCCR)*. IEEE, 2021, pp. 267–271.
- [36] L. Yu, E. Yang, and B. Yang, "Afe-orb-slam: robust monocular vslam based on adaptive fast threshold and image enhancement for complex lighting environments," *Journal of Intelligent & Robotic Systems*, vol. 105, no. 2, pp. 1–14, 2022.
- [37] G. G. Scandaroli, M. Meillard, and R. Richa, "Improving ncc-based direct visual tracking," in *European conference on Computer Vision*. Springer, 2012, pp. 442–455.
- [38] A. Crivellaro and V. Lepetit, "Robust 3d tracking with descriptor fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3414–3421.
- [39] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 422–429, 2019.
- [40] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 4523–4530.
- [41] J. Huang and S. Liu, "Robust simultaneous localization and mapping in low-light environment," *Computer Animation and Virtual Worlds*, vol. 30, no. 3-4, p. e1895, 2019.
- [42] P. Kim, H. Lee, and H. J. Kim, "Autonomous flight with robust visual odometry under dynamic lighting conditions," *Autonomous Robots*, vol. 43, no. 6, pp. 1605–1622, 2019.
- [43] Z. Chen and C. Heckman, "Robust pose estimation based on normalized information distance," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2217–2223.
- [44] 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.
- [45] R. Pautrat, J.-T. Lin, V. Larsson, M. R. Oswald, and M. Pollefeys, "Sold2: Self-supervised occlusion-aware line description and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11368–11378.
- [46] R. Pautrat, D. Barath, V. Larsson, M. R. Oswald, and M. Pollefeys, "DeepLsd: Line segment detection and refinement with deep image gradients," *arXiv preprint arXiv:2212.07766*, 2022.
- [47] A. Bartoli and P. Sturm, "Structure-from-motion using lines: Representation, triangulation, and bundle adjustment," *Computer vision and image understanding*, vol. 100, no. 3, pp. 416–441, 2005.
- [48] M. Kasper, S. McGuire, and C. Heckman, "A benchmark for visual-inertial odometry systems employing onboard illumination," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5256–5263.
- [49] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visual-inertial slam using nonlinear optimization," *Proceedings of Robotics Science and Systems (RSS)* 2013, 2013.
- [50] I. Cvišić, I. Marković, and I. Petrović, "Soft2: Stereo visual odometry for road vehicles based on a point-to-epipolar-line metric," *IEEE Transactions on Robotics*, 2022.
- [51] R. Song, R. Zhu, Z. Xiao, and B. Yan, "Contextavo: Local context guided and refining poses for deep visual odometry," *Neurocomputing*, 2023.
- [52] S. Zhang, J. Zhang, and D. Tao, "Towards scale consistent monocular visual odometry by learning from the virtual world," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 5601–5607.
- [53] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardos, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.