DK-VINS: Robust Deep Visual-Inertial SLAM for Embedded Devices

1st Yibo Dou
Tongji University
Shanghai, China
douyibo@tongji.edu.cn

2rd Qin Zhu
Hunan University
Hunan, China
B251000750@hnu.edu.cn

***Abstract*—Deep learning-based local features offer great promise for improving the robustness of visual-inertial SLAM, yet their performance often degrades in textureless or motion-blurred scenes—particularly on resource-constrained platforms. To address this, we propose DK-VINS, a robust deep Visual-Inertial Simultaneous Localization and Mapping system that integrates the reinforcement learning-trained detector DISK and the graph neural network matcher LightGlue into the front-end of VINS-Fusion. On the most challenging sequences of the EuRoC dataset, DK-VINS achieves accuracy and robustness comparable to state-of-the-art VI-SLAM systems. By combining ONNX-based efficient inference with a DBoW3 loop closure dictionary built offline from DISK descriptors, DK-VINS runs efficiently on embedded Jetson platforms. Experiments demonstrate that our learning-driven front-end significantly enhances VI-SLAM robustness in real-world degraded environments while maintaining real-time performance.**

***Index Terms*—Visual inertial simultaneous localization and mapping, Deep local features, Reinforcement learning, Graph neural networks**

I. Introduction

In practical visual-inertial SLAM (VI-SLAM) deployments—especially on resource-constrained platforms such as drones or mobile robots, front-end reliability under perceptually degraded conditions remains a critical challenge. While tightly coupled systems like VINS-Mono [1] and VINS-Fusion [2] perform well in typical scenes, their handcrafted features—e.g., FAST+BRISK [3], [4] or Shi-Tomasi [5]—often fail in textureless or motion-blurred frames, causing feature dropout or front-end collapse.For instance, on the EuRoC dataset [6], sequences such as *V2_03_difficult* (low illumination) and *MH_04_difficult* (aggressive motion) frequently lead to tracking loss in conventional pipelines due to insufficient or unstable keypoints.

Although learning-based local features excel in image matching, their integration into VI-SLAM faces two challenges: (1) *Objective misalignment*: Features like SuperPoint [7] and ALIKED [8] are trained on generic images without optimizing for geometric task success; (2) *Integration complexity*: Many deep VIO methods either use end-to-end pose regression (e.g., DeepVIO [9]), sacrificing geometric consistency, or require custom back-ends, hurting reproducibility and increasing engineering cost.

DISK [10] addresses the first issue by using policy gradient reinforcement learning to directly optimize features for geometric tasks, achieving state-of-the-art performance on low-texture benchmarks. Inspired by this, we integrate DISK with LightGlue into an industrial-grade VI-SLAM pipeline, presenting DK-VINS(Deep Keypoint Visual-Inertial SLAM)—a robust deep Visual-Inertial Odometry for embedded devices.DK-VINS is built upon the framework of VINS-Fusion. Through the collaborative design of three modules, it achieves the integration of accuracy and efficiency: (1) DISK feature extractor - generating high-repetition and geometrically consistent key points in areas with scarce information; (2) LightGlue matcher - using graph neural networks to output high-confidence and high-inlier-rate sub-pixel matches without the requiring for RANSAC post-processing; (3) DBoW3 loop closure detector - binarizing the DISK descriptors to construct a bag of words, achieving efficient loop closure. As shown in Fig.1, DK-VINS unifies accuracy and efficiency through three co-designed modules.Crucially, DK-VINS preserves the geometric integrity of classical optimization, ensuring metric accuracy for real-world robotics applications.

The main contributions of this work are:

- We propose DK-VINS—a plug-and-play front-end that integrates the reinforcement learning–trained DISK detector and graph-based LightGlue matcher into an industrial VI-SLAM framework. DISK optimizes keypoint placement via policy gradients for geometric success, while LightGlue enables high-fidelity matches through contextual reasoning, greatly improving robustness in low-texture or motion-blurred scenes.
- We achieve seamless integration of deep features with classical VI-SLAM by building an offline DBoW3 vocabulary from DISK descriptors for loop closure. Without altering core back-end components (e.g., IMU preintegration or marginalization), DK-VINS attains state-of-the-art accuracy and long-term consistency, proving that learning-based front-ends can enhance real-world VI-SLAM without sacrificing geometric integrity.
- We design a lightweight edge deployment pipeline using asynchronous ONNX [11] inference. DK-VINS runs at 38FPS (26ms/frame) on Jetson Orin NX, demonstrating that RL- and GNN-based front-ends are practical for resource-constrained platforms—offering high accuracy, robustness, and engineering feasibility.

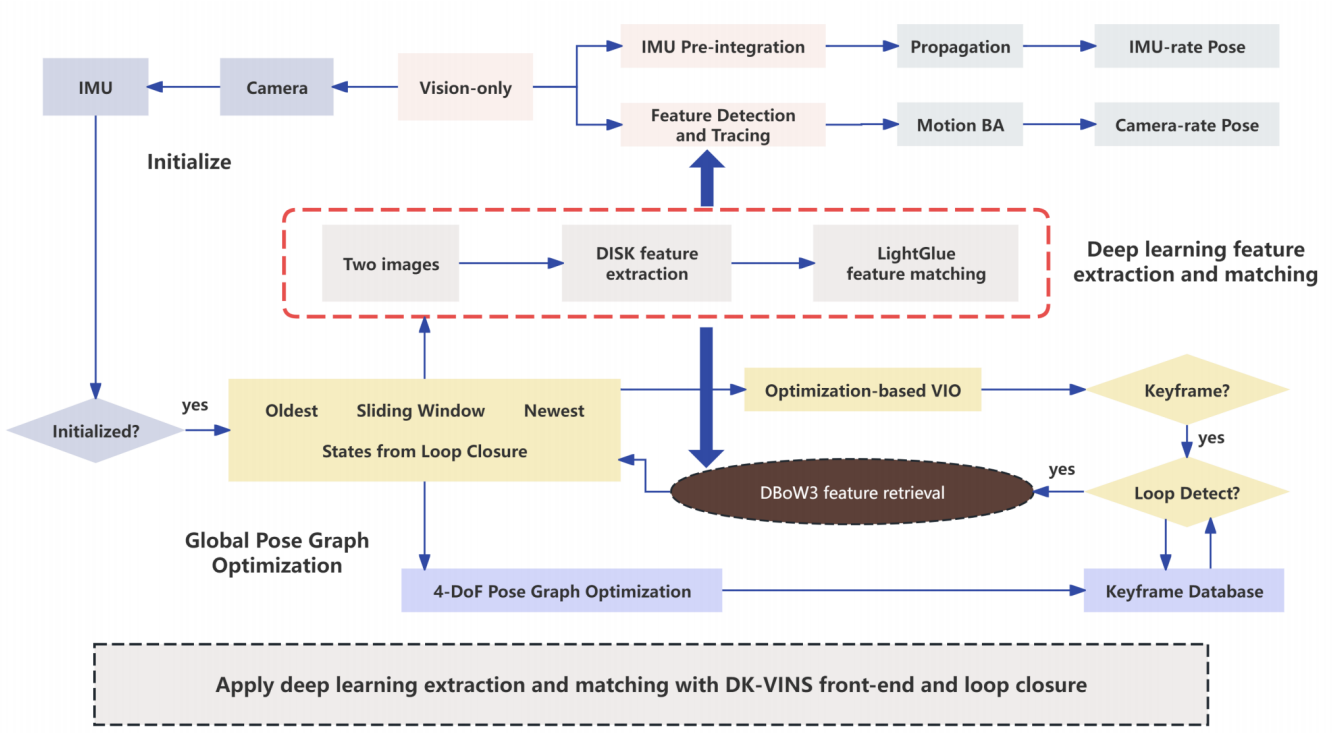


Fig. 1. DK-VINS framework. The system fuses camera and IMU data, employing DISK for keypoint detection and LightGlue for high-precision matching in the front-end. IMU preintegration and visual observations jointly drive sliding-window optimization. Loop closure is detected using DBoW3 with an offline-built DISK descriptor vocabulary, triggering 4-DoF pose-graph optimization to achieve robust and efficient embedded VI-SLAM.

II. Proposed Methods

DK-VINS is built upon the tightly coupled stereo-inertial framework of VINS-Fusion. Inspired by the front-end enhancement strategy of SuperVINS [12] and the VIO scheduling approach of EdgeVIO [13], DK-VINS retains the core back-end mechanisms of VINS-Fusion—including IMU preintegration [14], sliding-window optimization, and marginalization—while integrating three key components into a plug-and-play, task-oriented front-end architecture: a reinforcement learning-trained feature extractor (DISK), a graph neural network-based matcher (LightGlue), and a DBoW3 loop closure vocabulary constructed from DISK descriptors. This co-designed front-end effectively bridges the gap between generic deep features and the geometric requirements of VIO. The core innovation of DK-VINS lies in the synergistic integration of these three components:

A. Front-End: DISK Feature Extraction

To address the fragility of handcrafted features in perceptually degraded environments, we adopt DISK as our front-end feature extractor. DISK is a deep learning-based detector and descriptor trained end-to-end via reinforcement learning to directly maximize the success of downstream geometric tasks.yielding uniformly distributed and highly repeatable keypoints—even in information-scarce regions like dark areas, blank walls, or motion-blurred zones.

Given an input image I , the DISK network (a modified U-Net) outputs two dense maps: a keypoint heatmap $\mathcal{K} \in \mathbb{R}^{H\times W}$ and a descriptor map $\mathcal{D} \in \mathbb{R}^{H\times W\times 128}$. During training, keypoints are sampled probabilistically from $\mathcal{K}$. The image is divided into a grid of cells of size $h \times h$ (we use $h = 8$). Within each cell u , a location p is selected with probability:

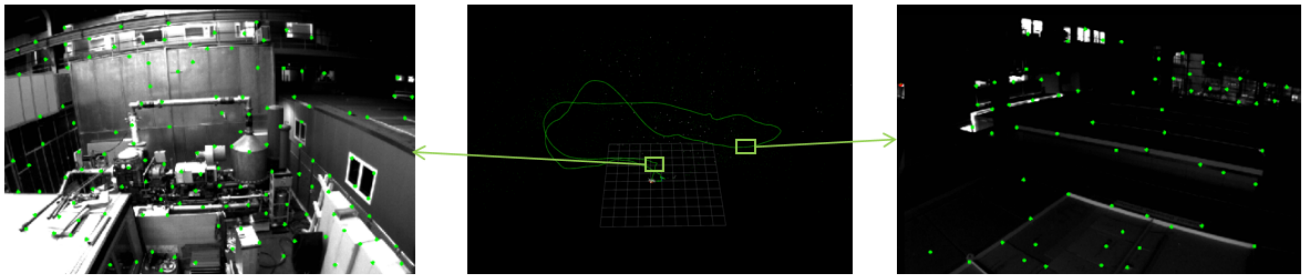
$$P(p|\mathcal{K}_u) = \underbrace{\text{softmax}(\mathcal{K}_u)_p}_{\text{relative preference}} \cdot \underbrace{\sigma(\mathcal{K}_u(p))}_{\text{absolute quality}}, \tag{1}$$

where σ is the sigmoid function. This dual formulation allows the model to learn both the relative saliency of points within a local neighborhood and their absolute distinctiveness for matching.

The core innovation of DISK lies in its training objective. Instead of using proxy losses, it employs a policy gradient method that directly optimizes the expected reward from geometrically verified matches. For a pair of images (A, B) , a match between features i and j is assigned a reward $r(i \leftrightarrow j)$ based on ground-truth geometry (e.g., reprojection error):

$$r(i \leftrightarrow j) = \begin{cases} \lambda_{\text{tp}} & \text{if correct match,} \\ \lambda_{\text{fp}} & \text{if incorrect match,} \\ 0 & \text{if plausible but unverifiable.} \end{cases} \tag{2}$$

The network is then trained to maximize the expected total reward $\mathbb{E}[\sum r(i \leftrightarrow j)]$ via gradient ascent. This direct optimization ensures that the learned features are intrinsically



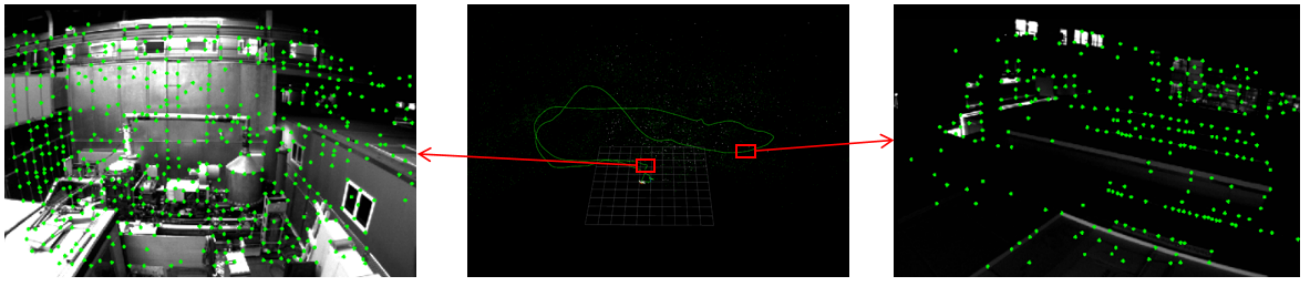


Fig. 2. VINS vs DK-VINS on MH04

geared towards producing high-quality, geometrically consistent correspondences.

At inference time, the probabilistic sampling is replaced by a deterministic process. We apply non-maximum suppression (NMS) on the heatmap $\mathcal{K}$ to obtain the final set of keypoints, which are more uniformly distributed than the grid-constrained training procedure. Their descriptors are L_2 -normalized vectors from $\mathcal{D}$. This design enables DISK to generate highly repeatable and discriminative keypoints even in textureless or blurred regions, providing a robust foundation for our VI-SLAM system.

Figure2 further showcases the performance difference between traditional geometric feature extractors and our learning-based frontend under typical degenerate conditions. In low-light scenarios, the number of discernible structural cues in the image drastically decreases, causing handcrafted features to fail in maintaining sufficient matchable points and stability.However, the learning-based front-end can still stably and accurately identify a large number of key points.

B. Matching: LightGlue with Geometric Validation

Given stereo keypoint sets extracted by DISK, we employ LightGlue [15] to establish robust feature correspondences. Let $\{(\mathbf{x}_i^L, \mathbf{d}_i^L)\}_{i=1}^{N_L}$ and $\{(\mathbf{x}_j^R, \mathbf{d}_j^R)\}_{j=1}^{N_R}$ denote the 2D positions and 128-D descriptors from the left and right images, respectively. Each node in the bipartite graph is initialized with a joint embedding:

$$\mathbf{h}_i^{(0)} = \left[\mathbf{W}_d \mathbf{d}_i;\ \mathbf{W}_p \hat{\mathbf{x}}_i\right], \tag{3}$$

where $\hat{\mathbf{x}}_i = 2(\mathbf{x}_i/\mathbf{s}) - 1$ normalizes coordinates to $[-1, 1]$ (with $\mathbf{s}$ the image size), and $\mathbf{W}_d$, $\mathbf{W}_p$ are learnable projection matrices.

LightGlue then performs cross-attention message passing between the two graphs. For a node i in the left image, its attention-weighted update from the right image is:

$$\mathbf{h}_i^{L,\text{new}} = \mathbf{h}_i^L + \sum_{j=1}^{N_R} \alpha_{ij} \cdot \mathbf{W}_v \mathbf{h}_j^R, \tag{4}$$

where the attention weight α_{ij} is computed as:

$$\alpha_{ij} = \frac{\exp\left((\mathbf{W}_q \mathbf{h}_i^L)^\top (\mathbf{W}_k \mathbf{h}_j^R)/\sqrt{d}\right)}{\sum_{j'} \exp\left((\mathbf{W}_q \mathbf{h}_i^L)^\top (\mathbf{W}_k \mathbf{h}_{j'}^R)/\sqrt{d}\right)}. \tag{5}$$

Here, $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are query, key, and value projection matrices, and d is the embedding dimension.

After L layers of such context-aware refinement, the final similarity score between a putative pair (i, j) is given by the dot product of their refined embeddings:

$$s_{ij} = \left(\mathbf{h}_i^{L,\text{final}}\right)^\top \mathbf{h}_j^{R,\text{final}}. \tag{6}$$

A soft assignment matrix $\mathbf{S} = [s_{ij}]$ is converted into hard matches via mutual nearest neighbor selection followed by geometric validation.

Crucially, thanks to LightGlue's contextual reasoning, the resulting matches exhibit high inlier ratios and sub-pixel precision—often eliminating the need for random sample consensus (RANSAC). Nevertheless, to ensure robustness, we apply a

lightweight epipolar geometry check using the known stereo calibration: a match $(\mathbf{x}_i^L, \mathbf{x}_j^R)$ is accepted only if

$$\left|(\mathbf{x}_j^R)^\top \mathbf{F} \mathbf{x}_i^L\right| < \tau, \tag{7}$$

where $\mathbf{F}$ is the fundamental matrix derived from calibrated camera intrinsics and baseline, and $\tau = 1.0$ is the reprojection tolerance in pixels. This hybrid strategy combines deep matching reliability with classical geometric rigor, forming a stable front-end for the VI-SLAM back-end.

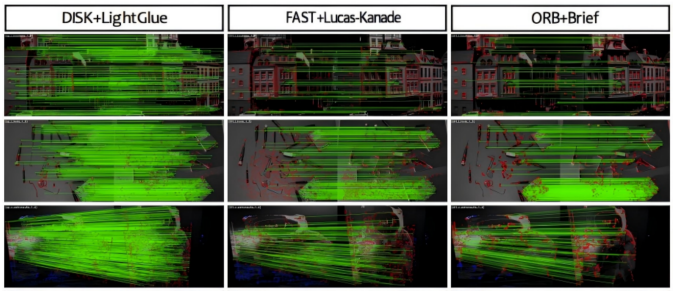


Fig. 3. This figure shows three feature extraction and matching methods for the same pair of images.

Feature matching is a cornerstone of visual-inertial odometry (VIO), directly influencing pose estimation accuracy and system efficiency. In this work, we adopt LightGlue as our matching backbone due to its lightweight architecture, deployment simplicity, and superior robustness. Unlike optical flow–based approaches, LightGlue establishes significantly more reliable correspondences—even among highly similar features—and maintains high matching quality under challenging conditions such as dynamic illumination and low light. Moreover, it can adaptively modulate its matching frequency based on scene complexity, enhancing real-time performance without compromising reliability. To highlight these advantages, we compare LightGlue against conventional geometric feature matchers; the results are shown in Fig.3.

Owing to LightGlue's geometry-aware design, DK-VINS achieves significantly higher inlier ratios compared to traditional nearest-neighbor matching with Lowe's ratio test, providing the back-end with more reliable visual constraints.

C. Back-End Integration and Deployment Optimization

Although LightGlue provides highly accurate feature matching, its computational overhead is too high to meet the demands of real-time global loop closure detection. To balance accuracy and efficiency, DK-VINS employs DBoW3 for scalable place recognition: it binarizes the floating-point DISK descriptors by thresholding each dimension at zero, converting them into compact binary representations. A visual vocabulary is built offline on the EuRoC dataset, and during online operation, DBoW3 [17] enables fast retrieval of loop closure candidate frames. Due to LightGlue's strong geometric consistency constraints, the resulting matches exhibit very low noise levels, allowing RANSAC post-processing to be omitted in most scenarios—significantly enhancing both the geometric consistency and computational efficiency of the back-end optimizer. Furthermore, the system implements an asynchronous ONNX inference pipeline to achieve real-time performance, ensuring low end-to-end latency with GPU acceleration.

III. Experiments and Results

In this section, we present an extensive experimental [18] evaluation of DK-VINS, including ablation studies and comparisons against state-of-the-art methods. The experiments are designed to validate the following aspects:

- Whether the proposed deep-feature front-end can significantly improve system robustness and accuracy under perceptually degraded conditions;
- Whether it is fully compatible with industrial-grade VI-SLAM pipelines, including loop closure detection;
- Whether it can be efficiently deployed on resource-constrained embedded platforms.

A. Dataset and Experimental Setup

To comprehensively evaluate the performance of DK-VINS, we conducted numerous experiments on the EuRoC MAV dataset [6], which is a widely used benchmark dataset for visual-inertial odometry (VIO). This dataset consists of a sequence of stereo images captured by a miniature aircraft equipped with a stereo camera and an inertial measurement unit (IMU), covering various challenging environments such as indoor offices and industrial settings. The experiment was conducted on a unified software and hardware platform and compared with the mainstream VI-SLAM methods. All algorithms were tested on a laptop with a NVIDIA RTX3050 GPU configured, Ubuntu 20.04 as the operating system, and ROS Noetic as the robot middleware. The system also had CUDA, CUDNN, and ONNX runtime configured to enable the inference of deep learning models. To verify the embedded deployment capability, the complete system was also run on the NVIDIA Jetson Orin NX (16GB version). For each sequence, we initialize the system using the data from the first few seconds to ensure a stable initialization. Then, we process the entire sequence frame by frame, recording the estimated trajectory and runtime performance. To ensure a fair comparison, all methods use the default parameter configuration, unless otherwise specified. To evaluate the robustness under conditions of degraded perception quality, we specifically analyzed sequences with significant motion blur (MH04) and low texture (V103). Additionally, we used DBoW3 for loop closure verification to prove the effectiveness of our position recognition module. Finally, we compared DK-VINS with the most advanced methods, including VINS-Fusion and ORB-SLAM3 [19], focusing on comparing their performance in handling complex scenes and whether they can maintain real-time capabilities.

B. System Comparison and Ablation Study

To systematically validate the effectiveness of the proposed method, we conduct ablation studies across all 11 EuRoC sequences and compare our approach against representative

TABLE I
RMSE (M) OF ATE ON EUROC DATASET

Model	MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203
Ours (Mono+IMU)	0.07	0.07	0.06	0.11	0.08	0.04	0.03	0.12	0.05	0.07	0.15
Ours (Stereo)	0.12	0.13	0.10	0.15	0.12	0.12	0.08	0.15	0.09	0.11	0.20
Ours (Stereo+IMU)	0.08	0.06	0.05	0.14	0.07	0.04	0.04	0.10	0.05	0.06	0.07
ORB-SLAM3 (Mono+IMU)	0.03	0.12	0.05	0.15	0.08	0.14	0.07	0.07	0.18	0.04	0.10
ORB-SLAM3 (Stereo+IMU)	0.06	0.10	0.09	0.13	0.16	0.12	0.04	0.11	0.05	0.08	0.06
VINS-Fusion (Mono+IMU)	0.09	0.11	0.08	0.13	0.15	0.05	0.07	0.19	0.055	0.095	0.25
VINS-Fusion (Stereo)	0.217	0.224	0.180	0.280	0.243	0.250	0.120	0.350	0.093	0.100	0.460
VINS-Fusion (Stereo+IMU)	0.170	0.155	0.130	0.290	0.290	0.080	0.230	0.320	0.070	0.095	0.100

TABLE II
RMSE (M) OF ATE ON EUROC DATASET FOR ABLATION STUDY

Model	MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203
Baseline VINS-Fusion (Mono+IMU)	0.090	0.110	0.080	0.130	0.150	0.050	0.070	0.190	0.055	0.095	0.250
+ DISK (NN matching)	0.080	0.090	0.070	0.120	0.120	0.040	0.050	0.150	0.050	0.080	0.200
+ LightGlue (orig. feat.)	0.100	0.100	0.080	0.140	0.140	0.060	0.060	0.170	0.060	0.090	0.220
Ours (DISK + LightGlue)	**0.070**	**0.070**	**0.060**	**0.110**	0.080	0.040	**0.030**	**0.120**	0.050	**0.070**	**0.150**
Baseline VINS-Fusion (Stereo)	0.217	0.224	0.180	0.280	0.243	0.250	0.120	0.350	0.093	0.100	0.460
+ DISK (NN matching)	0.160	0.170	0.140	0.200	0.180	0.180	0.090	0.250	0.080	0.090	0.300
+ LightGlue (orig. feat.)	0.180	0.190	0.150	0.220	0.200	0.200	0.100	0.280	0.090	0.100	0.350
Ours (DISK + LightGlue)	**0.120**	**0.130**	**0.100**	**0.150**	**0.120**	**0.120**	**0.080**	**0.150**	0.090	0.110	**0.200**
Baseline VINS-Fusion (Stereo+IMU)	0.170	0.155	0.130	0.290	0.290	0.080	0.230	0.320	0.070	0.095	0.090
+ DISK (NN matching)	0.120	0.110	0.100	0.200	0.180	0.060	0.150	0.250	0.060	0.070	0.070
+ LightGlue (orig. feat.)	0.140	0.130	0.110	0.220	0.200	0.070	0.180	0.280	0.060	0.080	0.080
Ours (DISK + LightGlue)	**0.080**	**0.060**	**0.050**	**0.140**	**0.070**	**0.040**	**0.040**	**0.100**	**0.050**	**0.060**	0.070

SLAM systems, including VINS-Fusion and ORB-SLAM3 [19]. All systems employ identical sensor calibration and initialization protocols to ensure a fair comparison [20].

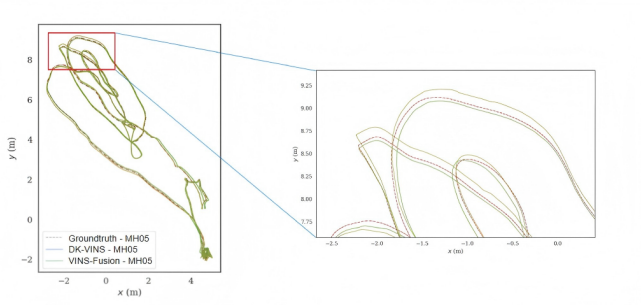


Fig. 4. Comparison of Trajectory Errors on the MH01 dataset

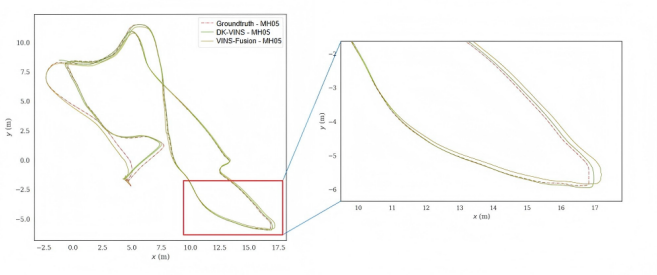


Fig. 5. Comparison of Trajectory Errors on the MH05 dataset

we perform statistical analysis and trajectory error visualization following the methodology in. The trajectory comparisons on the MH01 and MH05 datasets are presented in Figures4 and5, respectively, clearly demonstrating the superiority of DK-VINS.

Second, we quantitatively evaluate multiple algorithms using the Absolute Trajectory Error (ATE). As shown in TableI, under the stereo-inertial configuration, DK-VINS achieves the best or second-best accuracy across all 11 EuRoC sequences, with an average RMSE of 0.070 m. This significantly outperforms ORB-SLAM3 (stereo-inertial average RMSE: 0.110 m), reducing the trajectory error by 36.4%, and substantially improves upon VINS-Fusion (stereo-inertial average RMSE: 0.179 m) by 60.9To further verify the contribution of each proposed component, we present a detailed ablation study in TableII. Under the same evaluation protocol, the complete pipeline (DISK + LightGlue) consistently and significantly surpasses the VINS-Fusion baseline across all three sensor fusion modes: In the monocular-inertial mode, DK-VINS achieves an average RMSE of 0.089 m, improving by 27.6 over VINS-Fusion (0.123 m);In the stereo-only mode, it achieves an average RMSE of 0.151 m, yielding a 33.5 improvement over VINS-Fusion (0.227 m); In the stereo-inertial mode, it achieves an average RMSE of 0.070 m, representing a 60.9 reduction in error compared to VINS-Fusion (0.179 m). These results conclusively demonstrate that the synergistic design of DISK and LightGlue delivers consistent and substantial performance gains across diverse sensor configurations.

C. Embedded Deployment Performance

We evaluate the end-to-end performance of DK-VINS—including front-end processing, back-end

optimization, and loop closure—on an NVIDIA Jetson Orin NX (16 GB RAM), with all deep learning models executed via ONNX Runtime. The system achieves 38 FPS (26.3 ms per frame) on this embedded platform, with an average memory footprint of 330 MB (peak ¡ 390 MB). On a desktop platform equipped with an RTX 3050 GPU running Ubuntu 20.04, DK-VINS further reaches 83 FPS (12.1 ms/frame), and maintains 56 FPS (17.8 ms/frame) in CPU-only mode. These results demonstrate that DK-VINS not only delivers high accuracy but also meets the real-time requirements of typical visual-inertial (VI) sensor streams across diverse hardware configurations.the results are shown in TableIII.

TABLE III
RUNTIME BREAKDOWN ON JETSON ORIN NX (MS PER FRAME)

Component	Time (ms)	% of Total
DISK Feature Extraction	8.2	31.2
LightGlue Matching	5.9	22.4
Deep Front-End	**14.1**	**53.6**
IMU Preintegration	2.1	8.0
Sliding-Window Optimization	7.5	28.5
Marginalization	1.2	4.6
Back-End	**10.8**	**41.1**
Loop Closure (DBoW3)	1.4	5.3
Total	**26.3**	**100.0**

IV. CONCLUSION

We present DK-VINS—a robust deep VI-SLAM system tailored for real-time embedded operation in perceptually degraded real-world environments. By re-engineering VINS-Fusion's front-end, it seamlessly integrates task-oriented feature extraction (DISK), context-aware matching (LightGlue), and a classical VI-SLAM back-end—achieving strong robustness, geometric accuracy, and real-time performance without end-to-end pose regression or learned loop closure.Thanks to ONNX export and an asynchronous inference pipeline, it runs efficiently on embedded platforms, demonstrating practical viability for resource-constrained robots. Despite its strengths, DK-VINS has limitations. Future work will develop adaptive keypoint selection to balance accuracy and efficiency under tight computational budgets.

REFERENCES

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

[2] T. Qin, J. Pan, S. Cao, and S. Shen, "VINS-Fusion: Tightly-coupled visual-inertial fusion for state estimation," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), Montreal, QC, Canada, 2019, pp. 5971-5978.

[3] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in Proc. Eur. Conf. Comput. Vis. (ECCV), Graz, Austria, 2006, pp. 430-443.

[4] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Barcelona, Spain, 2011, pp. 2548-2555.

[5] J. Shi and C. Tomasi, "Good features to track," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 1994, pp. 593-600.

[6] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Wirth, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," Int. J. Robot. Res., vol. 35, no. 10, pp. 1157-1163, Sep. 2016.

[7] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 224-236.

[8] J. Yang, Z. Chen, S. Li, and J. Lai, "ALIKED: Keypoints with adaptive localization and invariance for efficient description," in Proc. Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, 2023.

[9] H. Zhang, M. Liu, and S. Shen, "DeepVIO: End-to-end visual-inertial odometry with deep learning," IEEE Robot. Autom. Lett., vol. 8, no. 6, pp. 3451-3458, Jun. 2023.

[10] M. Tyszkiewicz, P. Fua, and T. Trzcinski, "DISK: Learning local features with policy gradient," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, 2023, pp. 12278-12288.

[11] J. Bai, X. Lu, K. Zhang, Y. Wang, Y. Liu, and Y. Shi, "ONNX Runtime: A universal inference engine for machine learning models," in Proc. ACM SIGPLAN Int. Conf. Mach. Learn. Syst. (MLSys), Boston, MA, USA, 2023, pp. 112-125.

[12] X. Lu, Y. Zhang, L. Wang, et al., "SuperVINS: Deep feature-based visual-inertial navigation system," IEEE Sens. J., vol. 24, no. 5, pp. 7890-7901, Mar. 2024.

[13] Y. Wu, Z. Wu, and J. Zhang, "EdgeVIO: Efficient visual-inertial odometry on edge devices with adaptive feature scheduling," IEEE Robot. Autom. Lett., vol. 8, no. 9, pp. 5671-5678, Sep. 2023.

[14] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," IEEE Trans. Robot., vol. 33, no. 1, pp. 1-22, Feb. 2017.

[15] P. Edgar, T. Trzcinski, P. Fua, and M. Tyszkiewicz, "LightGlue: Local feature matching at light speed," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Paris, France, 2023, pp. 2259-2269.

[16] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "XFeat: Accelerated features for lightweight image matching," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, 2024, pp. 2682-2691.

[17] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," IEEE Trans. Robot., vol. 28, no. 5, pp. 1188-1197, Oct. 2012.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol. 24, no. 6, pp. 381-395, Jun. 1981.

[19] C. Campos, R. Elvira, J. J. Gómez-Olmedo, J. M. González-Mora, and J. Civera, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," IEEE Trans. Robot., vol. 37, no. 6, pp. 1874-1890, Dec. 2021.

[20] S. Yang, S. A. Scherer, and A. Zell, "Robust visual-inertial odometry using inlier-ratio-aware feature management," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Kyoto, Japan, 2022, pp. 8892-8899.