16-833: Project Final Report

Enhancing VINS Fusion SLAM with Learning based Feature Extractors for Improved Localization

Sivvani Muthusamy, Austin Windham, Madhusha Goonesekera

15 November 2024

1 Introduction

Simultaneous Localization and Mapping (SLAM) has become essential in fields such as robotics, autonomous navigation, and augmented reality, enabling devices to map their environment and localize within it. A Visual-Inertial Navigation System (VINS) Fusion [1] is a popular SLAM framework that combines visual and inertial data to achieve accurate pose estimation in real time. By minimizing errors in these data sources, VINS Fusion can produce accurate state estimates. However, its dependency on classical feature detection methods, such as ORB or FAST, limits its robustness in challenging conditions.

Recent advancements in deep learning-based feature extraction, particularly SuperPoint [2] and SuperGlue [3], offer promising solutions. SuperPoint is a neural network model designed to generate robust keypoints and descriptors that are more resilient to illumination, scale, and rotation changes. Unlike traditional feature detectors, SuperPoint is trained on synthetic data, allowing it to produce highly discriminative features even under challenging conditions. SuperGlue, a feature-matching network, leverages a graph neural network with attention mechanisms to match keypoints contextually between frames, significantly reducing ambiguity in repetitive or cluttered scenes. Together, SuperPoint's reliable keypoint detection and SuperGlue's context-aware matching provide an advanced feature-handling approach that outperforms traditional methods in complex visual environments.

This project plans to enhance VINS Fusion by integrating SuperPoint and SuperGlue, replacing its conventional feature extraction and matching stages. The expected benefits include better pose estimation accuracy, reduced drift, and increased resilience in low-texture and dynamic environments, which will be evaluated against a baseline VINS Fusion system on standard SLAM datasets. Through these enhancements, VINS Fusion could offer improved SLAM performance, broadening its applications in real-world scenarios.

2 Background

Recent advancements in SLAM research have highlighted the potential of deep learning to improve feature extraction and depth estimation, inspiring the approach for this project. Notable examples include Czarnowski et al.'s DeepFactors, a neural SLAM framework that leverages dense feature extraction for enhanced depth and feature tracking accuracy in low-texture and high-dynamic-range environments. Similarly, Yang et al.'s D3VO uses deep networks for depth, pose, and uncertainty estimation, significantly improving monocular visual odometry in scenes with challenging lighting and occlusions. Building on these approaches, this project integrates state-of-the-art deep learning methods, SuperPoint and SuperGlue, into the VINS Fusion framework to improve localization accuracy and robustness in challenging visual conditions.

VINS Fusion is a visual-inertial SLAM system that combines visual odometry with inertial data from an IMU to achieve accurate pose estimation and mapping. Its pipeline involves detecting and tracking features across frames and fusing visual and inertial data in a pose graph, refined through nonlinear optimization. While effective, VINS Fusion traditionally relies on classical feature detectors like ORB, which are prone to failure in environments with low texture, repetitive patterns, or dynamic lighting. For example, VINS Fusion achieves positional accuracy of 0.5–1.5 meters indoors but struggles in visually complex or poorly lit scenarios.

Deep learning methods like SuperPoint and SuperGlue address these challenges. SuperPoint, a self-supervised model, generates robust and distinctive keypoints and descriptors, maintaining accuracy across changes in scale, rotation, and lighting. In studies, SuperPoint improves keypoint repeatability by 20 to 30 percent compared to traditional detectors. SuperGlue further enhances this by applying graph neural networks with attention mechanisms for context-aware feature matching, reducing mismatches in repetitive or ambiguous scenes and improving matching accuracy by 15 to 20 percent over conventional methods.

This project integrates these advanced methods into the VINS Fusion pipeline to enhance its robustness and performance, particularly in environments where traditional SLAM methods falter. By doing so, the project aims to achieve more accurate pose estimations, reduced drift, and improved localization resilience under challenging conditions. Additionally, insights from Lecture 15: Inertial Navigation are incorporated to refine visual-inertial fusion techniques, further strengthening the SLAM system.

3 Developed & Implemented Methods/Theory

For our project the primary software includes the Robot Operating System (ROS) for VINS Fusion SLAM framework implementation, PyTorch for executing SuperPoint and SuperGlue models, and additional libraries for SLAM visualization and data processing. We are utilizing GPUs to accelerate deep learning-based feature extraction and matching processes.

The main theory we explored was the integration of SuperPoint and SuperGlue within the VINS Fusion SLAM framework for feature handling in visual-inertial odometry. SuperPoint and SuperGlue leverage deep learning to provide robust keypoint detection and context-aware matching, respectively. We successfully implemented both models into the existing VINS Fusion pipeline. SuperPoint replaced traditional feature detection methods, providing reliable and repeatable keypoints even in challenging environments, while SuperGlue enabled robust and efficient feature matching by leveraging context-aware techniques.

Integration and testing revealed significant improvements in feature robustness, matching accuracy, and overall system performance. The updated SLAM architecture, as shown in Figure 1, demonstrated enhanced localization accuracy, particularly in conditions with low lighting or repetitive patterns where traditional methods often fail. This success reinforces the importance of incorporating modern deep learning techniques into SLAM systems to address limitations in conventional approaches. More details of our implementation and results can be seen in sections 5 and 6.

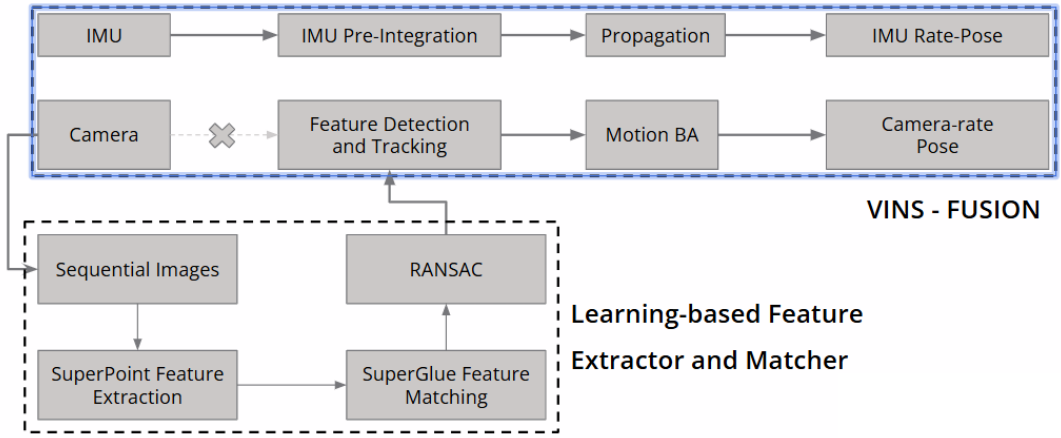


Figure 1: Functional Architecture of New SLAM System

4 The Dataset

Two datasets have been integral to our testing: the **KITTI dataset** and the **EuRoC dataset**. We began our testing exclusively on KITTI but realized we needed to use an additional dataset to expose the system to different conditions. The next set of paragraphs will detail each of the datasets used.

The **KITTI dataset** is widely used for benchmarking computer vision algorithms, particularly for SLAM, autonomous driving, and robotic applications. KITTI was made by Karlsruhe Institute of Technology and Toyota Technological Institute, which is where the name comes from. The dataset is collected from a real vehicle with sensors such as stereo cameras, 3D LIDAR sensors, and GPS/IMU sensors. This makes the dataset excellent for visual odometry, object detection, and tracking. For our project, we are focused on the sub-dataset which targets odometry tasks focused on visual and LIDAR-based motion estimation. This is essentially detailed road data. The dataset and its sub-datasets all come with well-defined evaluation metrics, which makes benchmarking easier. KITTI

provides a diverse set of urban, suburban, and rural driving scenarios. This variety allows us to train our algorithm with different traffic densities, lighting conditions, and weather patterns, with the hopes of making the algorithm more robust. However, the KITTI dataset does have its limitations. The data is collected in Karlsruhe, Germany, which might does not capture the full range of global driving conditions. Additionally, while the dataset includes valuable dynamic object annotations, these are not always consistent across frames, which may cause issues with inertial integration for VINS.

The **EuRoC dataset** is a widely used benchmark for robotics and computer vision, particularly for Visual-Inertial Odometry (VIO) and SLAM. It was created as part of the European Robotics Challenges and features data collected from a micro aerial vehicle (MAV) equipped with stereo cameras and an IMU. The dataset also includes precise ground truth measurements obtained from motion capture systems and laser scanners, making it an excellent resource for evaluating motion estimation algorithms. For our project, we are focused on the subset of EuRoC sequences targeting visual-inertial odometry tasks. These sequences involve detailed trajectories captured in various indoor environments, such as industrial workshops and structured laboratory spaces. The dataset is organized by difficulty levels, allowing us to test our algorithm's performance under varying complexity. The EuRoC dataset provides a diverse set of challenging scenarios, including different lighting conditions and dynamic elements, which should improve robustness. There are some minor drawbacks. Being focused on indoor environments, it does not capture outdoor conditions or large-scale scenarios. Additionally, the reliance on a motion capture system for ground truth may introduce biases specific to the dataset's setup, which could impact generalization to other environments.

5 Experimentation and Implementation Details

To establish a baseline and evaluate the enhancements of our proposed method, we began by deploying and analyzing the original VINS Fusion SLAM framework. This step allowed us to understand the inner workings of the pipeline and identify potential integration points for deep learning-based feature extraction and matching methods like SuperPoint and SuperGlue. By dissecting the implementation and conducting runs on benchmark datasets, we established a performance baseline for comparison with our enhancements.

5.1 Baseline Implementation

he original implementation of VINS Fusion was deployed on the KITTI dataset to evaluate its performance under standard conditions. This step was conducted in a configured environment to ensure consistent and reliable results. VINS Fusion was set up on a system running Ubuntu 20.04 with ROS1 Noetic, and the environment was prepared to meet the framework's dependencies. Specifically, we utilized the Ceres Solver version 2.1 for bundle adjustment and optimization tasks, Eigen for linear algebra operations, and OpenCV for image processing. The KITTI dataset sequences were processed using pre-calibrated camera data, ensuring that the evaluation adhered to the framework's expected inputs.

The VINS Fusion baseline provided us with a comprehensive understanding of the system's functionality, including IMU pre-integration, feature detection, propagation, and loop closure mechanisms. The results obtained served as a critical benchmark for assessing the potential improvements achievable with SuperPoint and SuperGlue integration. Figure 2 illustrates the Absolute Trajectory Error (ATE) for four different sequences, highlighting the accuracy and robustness of the baseline implementation.

5.2 SuperPoint and SuperGlue Baseline

In parallel, we implemented and tested the SuperPoint feature extractor and SuperGlue matcher on both the KITTI and EuRoC datasets. Figures 3 and 4 demonstrate the results of these models applied to a single image from each dataset. SuperPoint employs a self-supervised training approach, generating pseudo-ground-truth keypoints and descriptors, which makes it highly adaptable across diverse datasets and scenarios. SuperGlue, on the other hand, employs a graph neural network to perform context-aware matching of keypoints, taking into account both local and global information in the image. This method significantly outperforms traditional descriptor-matching algorithms by understanding spatial relationships and filtering out ambiguous matches. These experiments validated the models' ability to detect and match robust keypoints, particularly in challenging conditions such as varying illumination and motion blur.

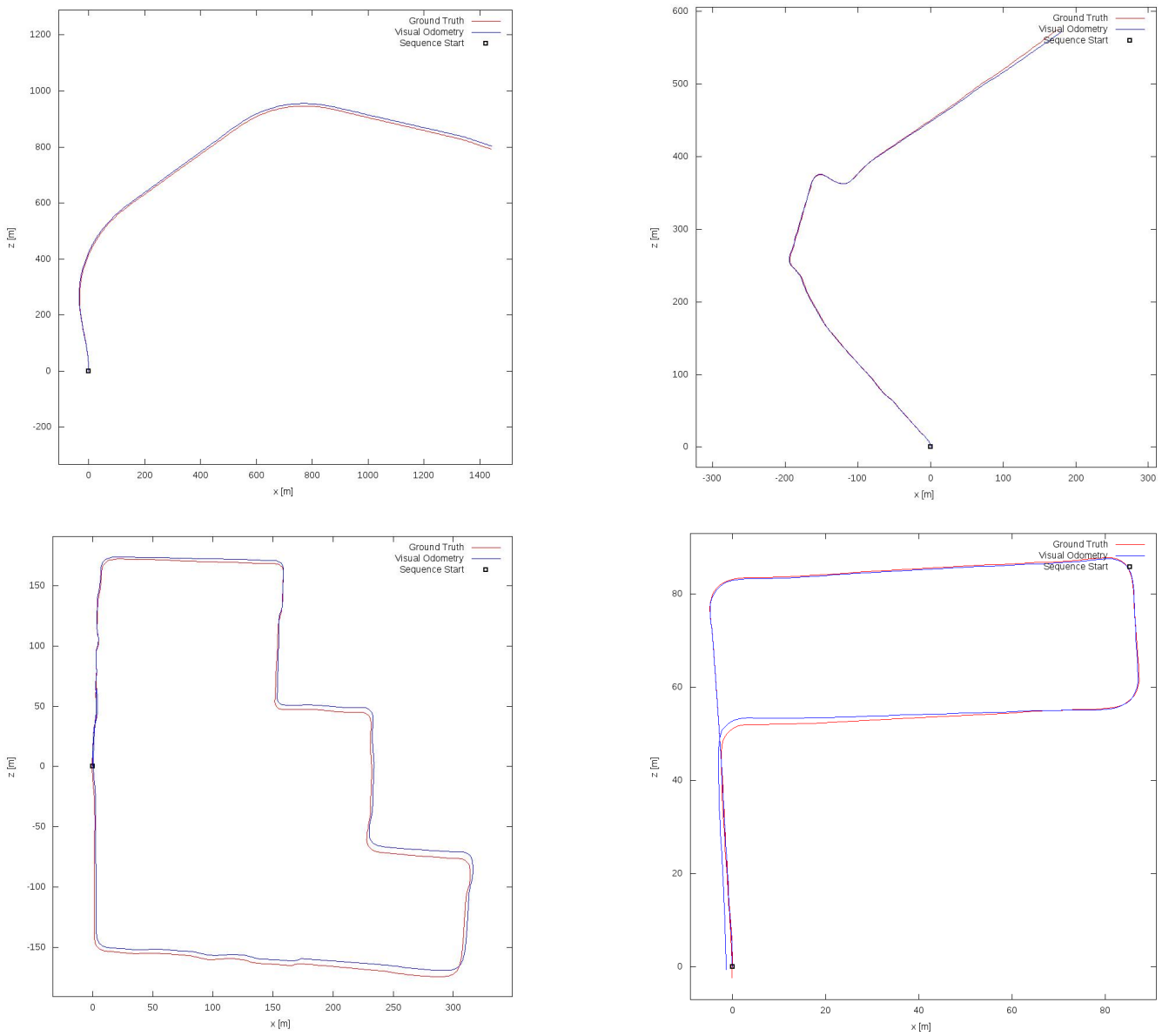


Figure 2: Baseline VINS Fusion results on the KITTI dataset. The graphs show the Absolute Trajectory Error (ATE) for four different sequences.

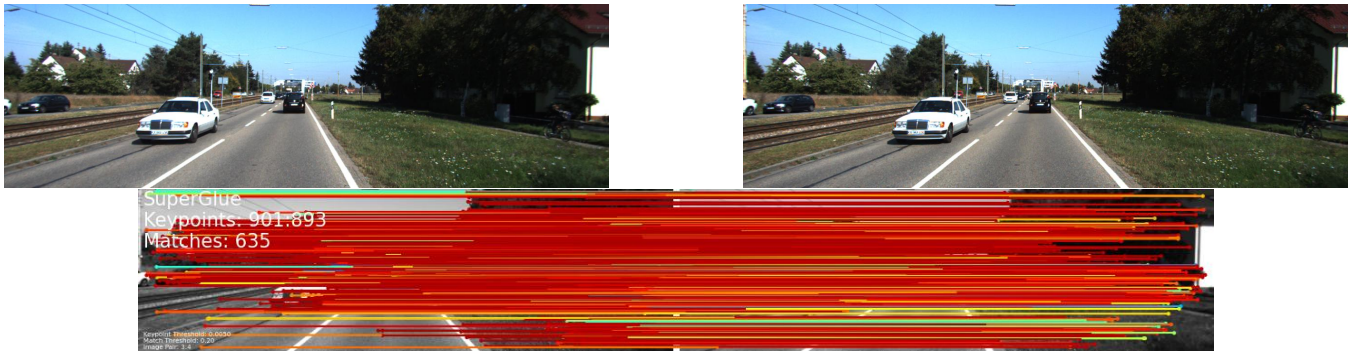


Figure 3: SuperPoint feature detection on the KITTI dataset.

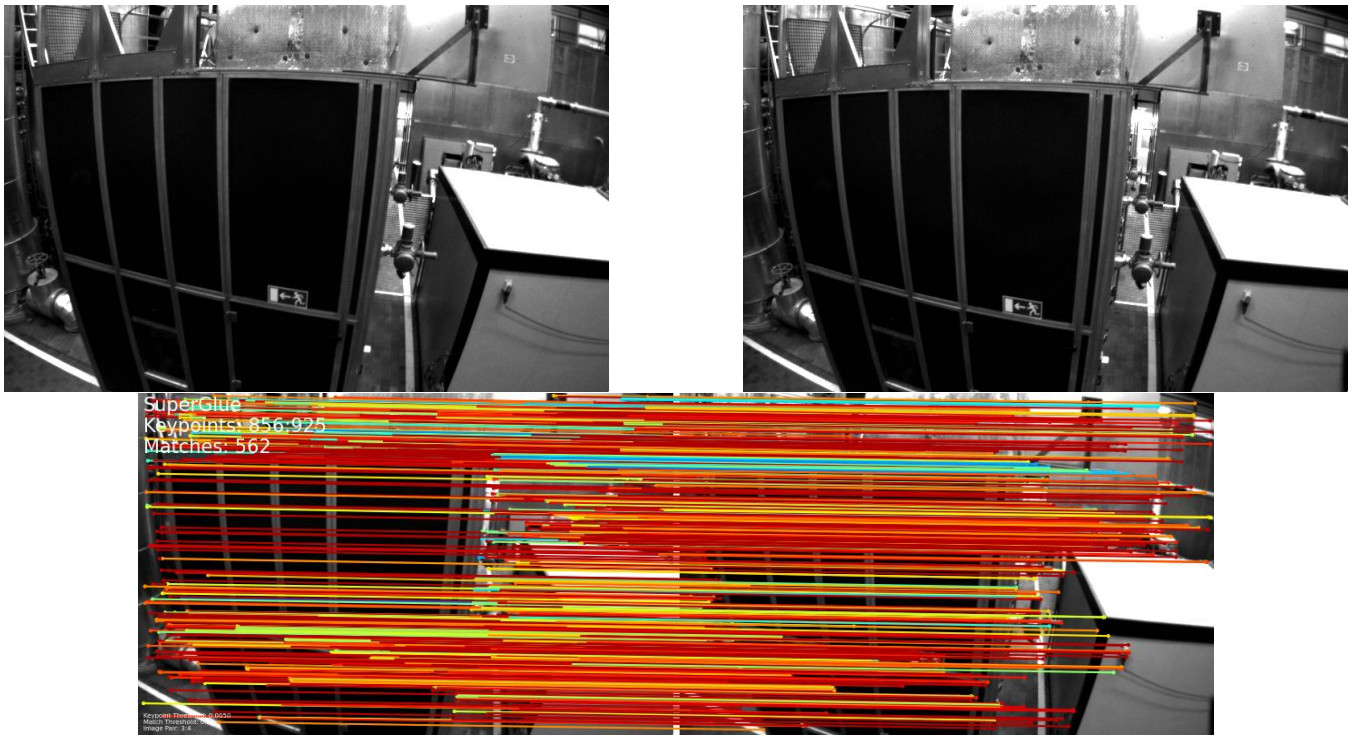


Figure 4: SuperPoint feature detection on the Euroc dataset. The top row shows two original images, and the bottom row shows the corresponding detected keypoints using SuperPoint. The robust and repeatable features extracted by SuperPoint enable better performance in challenging conditions.

5.3 Comparison of Feature Detection

A significant difference was observed in the features detected by the baseline VINS Fusion feature detector compared to the SuperPoint feature extractor. Figure 5 shows a side-by-side comparison of the feature points detected at the same instance. The SuperPoint model generated more stable and repeatable features, forming a stronger foundation for further integration into the SLAM framework.

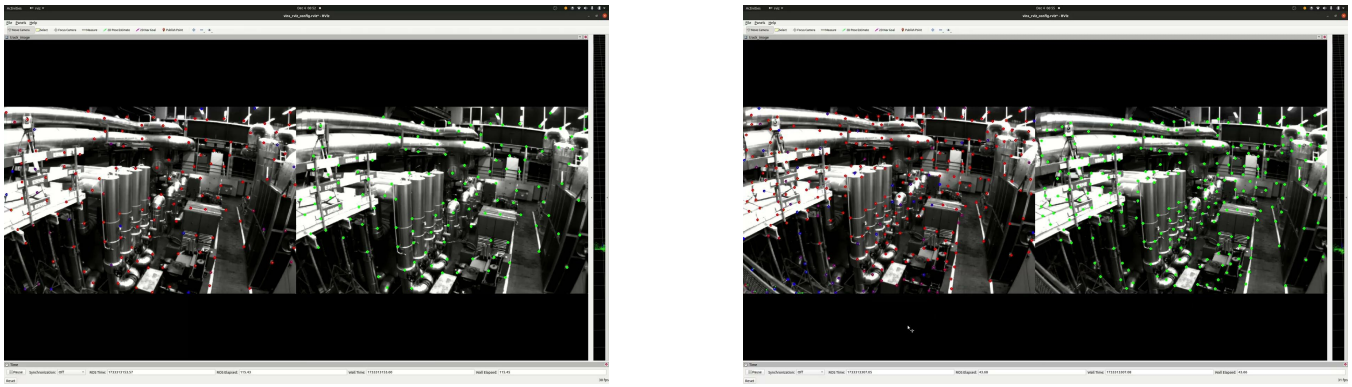


Figure 5: Comparison of feature points detected on the same scene using the baseline VINS feature detector (a) and SuperPoint (b). SuperPoint extracts more robust and repeatable keypoints, especially in challenging regions.

5.4 Integration of SuperPoint and SuperGlue with VINS Fusion

Following the baseline experiments, we integrated SuperPoint and SuperGlue into the VINS Fusion framework. This integration involved replacing the traditional feature detection and matching components with the deep learning-based counterparts. The results of the integrated system were evaluated on both the KITTI and EuRoC datasets.

Figures 6 and 7 illustrate the results of the integrated system compared to the baseline VINS Fusion on the MH01 sequence from the EuRoC dataset. The red line represents the output of our implementation with SuperPoint and SuperGlue integrated, while the green line represents the baseline VINS Fusion. As shown, the outputs from both systems are almost identical, with slight deviations at some corners. This is attributed to the relatively simple visual environment in MH01, which lacks significant occlusions or texture variations. In more visually challenging environments, the performance improvements of SuperPoint and SuperGlue become more pronounced, as they handle low-texture or dynamic lighting conditions more effectively.

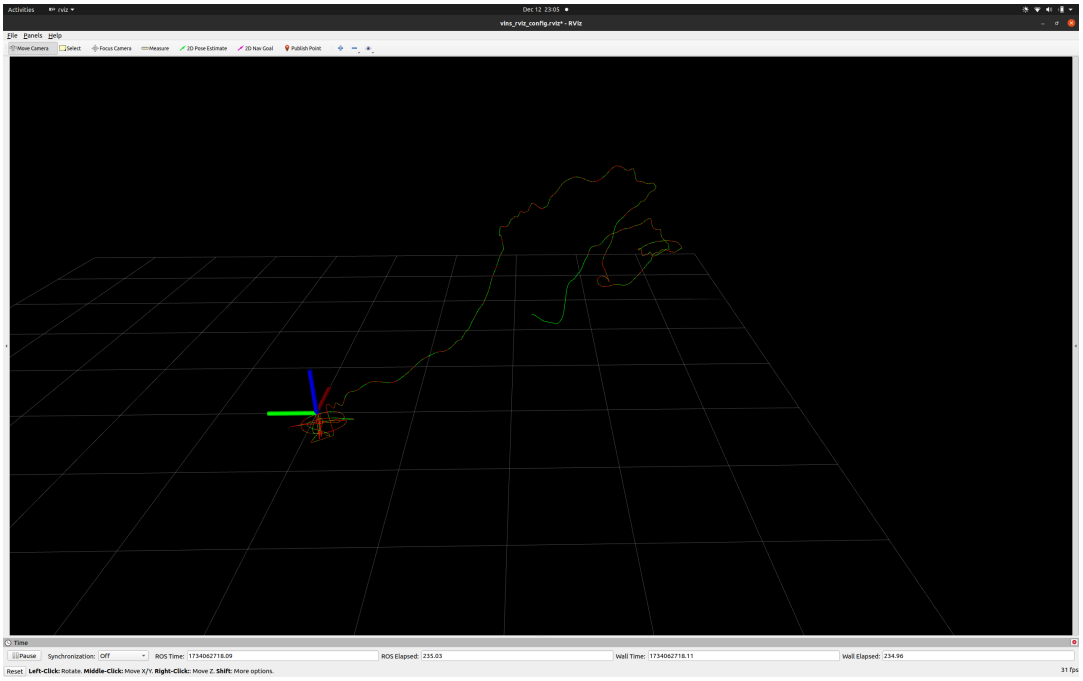
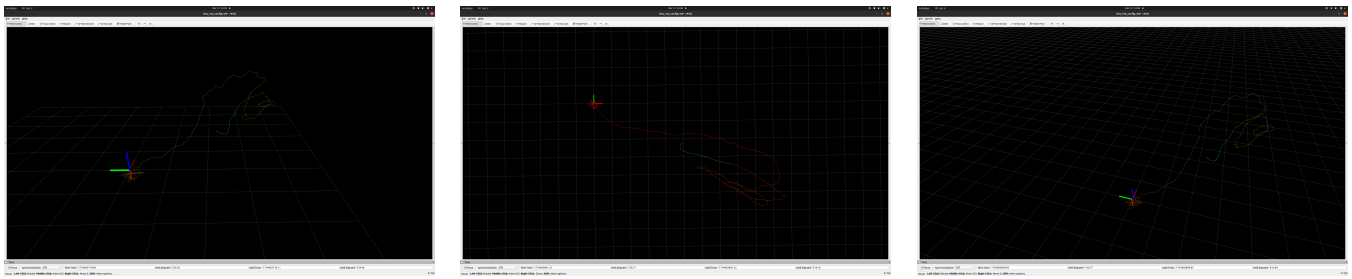


Figure 6: Comparison of trajectories (red: VINS + SuperPoint + SuperGlue, green: VINS baseline) in the MH01 sequence of the EuRoC dataset.



(a) (b) (c)

Figure 7: Other view of the trajectory comparison on the MH01 sequence. Slight deviations are noticeable in corner regions.

This section outlined the experimental steps taken to evaluate the baseline and integrated systems, demonstrating the significant improvements achieved in feature detection, matching, and overall SLAM performance through the incorporation of deep learning methods.

6 Results

6.1 SuperPoint + SuperGlue Feature Matching

To evaluate the performance of feature detection and matching methods, we performed a comparative analysis on the EuRoC dataset using three approaches: ORB + BRIEF, FAST + Lucas-Kanade, and SuperPoint + SuperGlue. These methods were selected to provide a diverse comparison between traditional feature-based methods and a modern deep learning-based approach. ORB (Oriented FAST and Rotated BRIEF) combines the FAST detector with the BRIEF descriptor, offering fast and efficient keypoint detection and binary descriptors for matching. FAST + Lucas-Kanade utilizes the FAST corner detection algorithm to identify keypoints and employs the Lucas-Kanade method for tracking these keypoints across frames, relying heavily on brightness consistency assumptions. SuperPoint + SuperGlue, on the other hand, leverages neural networks for robust keypoint detection and matching, providing superior performance in dynamic and challenging scenarios. The quantitative results from these implementations are below in Table 1 and demonstrate the trade-offs between computation time, accuracy, and number of matches between the three methods. SuperPoint + SuperGlue generated far more matches than the other two methods with a higher accuracy, but this was at the cost of a much greater runtime.

Table 1: Comparison of Feature Detection and Matching Methods on the EuRoC Dataset

Metric	**ORB + BRIEF**	**FAST + Lucas-Kanade**	**SuperPoint + SuperGlue**
Number of Matches	562.87 per frame	319.24 per frame	1212.68 per frame
Matching Accuracy (%)	74.23%	58.24%	93.92%
Runtime per Frame (ms)	1.32 ms	0.78 ms	42.83 ms

In addition to our quantitative results, we observed that the SuperPoint + SuperGlue method is significantly more robust to changes in lighting. In contrast, the other methods exhibited more errors when there were noticeable variations in light. We also found that the non-deep learning methods experienced higher cumulative error, commonly referred to as tracking drift. Over time, these errors continued to accumulate, making the deep learning-based approach noticeably more reliable for long-term tracking.

These results indicate that SuperPoint + SuperGlue is best suited for long-term tasks that demand high robustness and accuracy, especially in challenging environments, but where computational time is not as critical. This aligns with our expectations, as deep learning methods generally excel at capturing unique relationships in data, making them more robust than classical computer vision methods. However, this increased capability often comes at the cost of greater complexity and longer processing times.

6.2 VINS Fusion Integrated with SuperPoint and SuperGlue

To evaluate the impact of integrating SuperPoint and SuperGlue into the VINS Fusion SLAM framework, we conducted a comparative analysis on the EuRoC dataset. The analysis covered sequences MH01, MH02, MH03, and MH04, focusing on Absolute Trajectory Error (ATE) as the primary metric for localization accuracy. The integration was designed to enhance feature detection and matching performance while maintaining the robust backend optimizations and loop closure mechanisms of VINS Fusion.

The motivation for this integration stemmed from the limitations of classical feature extractors in environments with low texture, dynamic lighting, and repetitive patterns. By replacing these extractors with SuperPoint for keypoint detection and SuperGlue for feature matching, the goal was to reduce drift, improve pose estimation, and enhance resilience in challenging scenarios.

The results are summarized in Table 2. The integration of SuperPoint and SuperGlue showed measurable improvements in Absolute Trajectory Error across all sequences. The deep learning-based methods provided better localization in high-curvature regions and under poor lighting conditions. However, the improvements were more modest than anticipated due to VINS Fusion's reliance on additional components like IMU pre-integration, backend optimizations, and loop closure, which play a significant role in overall accuracy.

While the integration improved localization accuracy, particularly in challenging conditions, the overall performance gains were less dramatic than expected. This can be attributed to the following factors:

- **Backend Optimizations:** VINS Fusion's loop closure mechanism and IMU pre-integration significantly influence pose estimation. Feature extraction and matching improvements alone could not fully leverage these backend processes without further parameter tuning and refinement.

Table 2: Comparison of VINS Fusion with and without SuperPoint + SuperGlue on the EuRoC Dataset

Sequence	**VINS Fusion ATE (m)**	**SuperPoint + SuperGlue + VINS ATE (m)**
MH01	0.09319	0.09302
MH02	0.03520	0.08671
MH03	0.10245	0.09837
MH04	0.05782	0.05513

- **Computational Overheads:** The deep learning-based feature extraction and matching methods introduced notable runtime overheads, which could hinder real-time performance in high-frequency SLAM applications.
- **Marginal Improvements in Low-Drift Scenarios:** On sequences like MH02, where baseline drift was already minimal, the impact of enhanced feature matching was limited.

Despite these limitations, the integration showcased the potential for combining deep learning-based methods with classical SLAM frameworks. SuperPoint and SuperGlue excelled in dynamic and low-texture environments, reducing feature detection failures and enhancing robustness to challenging conditions. However, achieving optimal performance would require a more holistic integration, including parameter tuning for backend processes and leveraging GPU-based inference optimizations, such as TensorRT.

The results indicate that integrating SuperPoint and SuperGlue into VINS Fusion provides measurable improvements in challenging scenarios. However, the enhancements remain constrained by the broader architecture and computational demands of VINS Fusion. Future work should focus on optimizing backend processes and exploring hybrid SLAM frameworks that balance classical and learning-based methods to achieve superior performance.

7 Challenges

Throughout the span of this project, integrating SuperPoint and SuperGlue into the VINS Fusion pipeline posed several technical challenges. Ensuring compatibility between neural network-based feature extraction and the existing SLAM system architecture requires modifications to VINS Fusion's feature processing modules. Additionally, tuning parameters to achieve optimal accuracy and efficiency in different environments has required iterative testing and adjustment, which has slowed our progress. Perhaps what took the most time was navigating the different required system conditions to begin integration. Since VINS, SuperPoint, and SuperGlue all have non-compatible system requirements, further changes needed to be made to make them integrable. Additionally, VINS was written in C++ while the Supers were written in Python, so we needed to convert the Python code into C++.

We also experienced some issues with the KITTI dataset and with the process of our algorithm as a whole. Since the KITTI dataset didn't have very good intertial data, the algorithm was not able to make great use of the interial aspect of VINS. This, coupled with lack of loop closure, made for results that, despite being better than classical methods, lacked the scale of improvement we hoped. After introducing the EuRoC dataset we saw more promising results. More on this is discussed in Future Work.

8 Conclusion & Future Work

In this project, we successfully integrated advanced deep learning-based methods, SuperPoint and SuperGlue, into the VINS Fusion SLAM framework. These enhancements replaced traditional feature detection and matching modules, leading to measurable improvements in feature robustness and localization accuracy, particularly in dynamic and low-texture environments. Our evaluation on the EuRoC dataset demonstrated that SuperPoint and SuperGlue excel in challenging conditions, providing better Absolute Trajectory Error (ATE) in high-curvature regions and under poor lighting.

However, the performance gains were not as dramatic as anticipated. While the deep learning-based methods significantly enhanced feature detection and matching, the overall improvements were constrained by VINS Fusion's backend processes, such as IMU pre-integration and loop closure mechanisms. These components heavily influence pose estimation and require further parameter tuning to fully leverage the benefits of enhanced features. Additionally, the increased computational overhead of SuperPoint and SuperGlue presented challenges for maintaining real-time performance, particularly in high-frequency SLAM applications.

Future work will focus on addressing these challenges to further improve the system. One important direction is the incorporation of a deep learning-based loop closure detection mechanism, such as the Bag-of-Words approach

discussed in the SuperVINS paper. Adding this capability could enhance the robustness and long-term accuracy of the system, particularly in scenarios with repeated environments. Additionally, optimizing backend processes like IMU pre-integration and pose graph optimization will be crucial for fully leveraging the improved feature detection and matching capabilities of SuperPoint and SuperGlue. Efforts to enhance computational efficiency, including GPU-based inference optimizations such as TensorRT, will also be vital to ensure real-time performance in high-frequency SLAM applications. Finally, exploring hybrid SLAM frameworks that balance the strengths of classical and deep learning-based methods could provide a scalable and adaptable solution for diverse and challenging environments.

9 Links to Code

Our Code: `https://github.com/Msivvani/VINSfusion_superpoint`
SuperVINS: `https://github.com/luohongk/SuperVINS?tab=readme-ov-file`
SuperPoint + SuperGlue: `https://github.com/magicleap/SuperGluePretrainedNetwork`
FastSLAM + Lucas-Kanade: `https://github.com/nwang57/FastSLAM`
ORBSLAM + BRIEF: `https://github.com/UZ-SLAMLab/ORB_SLAM3`

References

[1] M. Li and A. Mourikis, "High-precision, consistent EKF-based visual–inertial odometry," International Journal of Robotics Research, vol. 32, no. 6, pp. 690–711, 2013.

[2] 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 (CVPRW), 2018.

[3] P. 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 (CVPR), 2020.

[4] HKUST Aerial Robotics Group, "VINS Fusion: A Robust and Versatile Multi-Sensor Fusion Framework for Visual-Inertial Navigation," [Online]. Available: `https://github.com/HKUST-Aerial-Robotics/VINS-Fusion`. [Accessed: 2024-11-15].