

A Robust Deep Learning-based Feature Extraction and Matching Technique for Visual Odometry in VSLAM

Madhu Sudan M P¹, Sudarshan Patil Kulkarni² and Aprameya C V³

¹⁻³JSS Science and Technology University Mysore 570006

Email: madhusudan.mp@jssstuniv.in

Abstract—Simultaneous Localization and Mapping or SLAM is a computational process that concurrently estimates an agent's location and builds a map of its unknown environment, utilizing approaches such as Light Detection and Ranging (LiDAR SLAM) and Visual SLAM (VSLAM), among which the VSLAM technique performs better for indoor localisation and mapping. Traditional VSLAM techniques like Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Oriented FAST and Rotated BRIEF (ORB) use geometric methods for feature extraction and matching. While these techniques are effective in static scenarios, they often struggle in dynamic or visually ambiguous environments due to their limitations in feature extraction and motion estimation. The current work addresses these challenges by incorporating deep learning techniques to improve VSLAM's reliability and adaptability in dynamic environment. Our approach leverages deep learning-based feature extractors and motion estimation models, designed to enhance the system's capability to interpret complex environments, thereby resulting in more robust and precise feature extraction and matching.

The deep learning-based model is designed using a custom-enhanced Densely-Connected Two-Stream Network Model (D2-Net Model) with regularisation and attention mechanisms. The conventional D2-Net architecture has been improved by incorporating additional convolution and dropout layers for feature extraction and dilated convolution layers with optimisation blocks to enhance feature discrimination, which makes the model more resilient to scene changes and viewpoint variations, especially for dynamic scenarios. Compared to traditional models, the developed architecture consumed lesser time for feature computation per image and better overall execution time for complete dataset. The designed model has been tested on the TUM-RGB-D fr1/xyz dataset, for which the model achieved 602 good feature matches per image pair with a high inlier ratio of 97.76% based on Random Sample Consensus (RANSAC) matching criteria, which showed strong consistency for the features extracted across all frames.

Keywords— VSLAM, feature extraction, feature matching, keypoint detection, model optimization.

I. INTRODUCTION

Simultaneous localization and mapping [SLAM] is a computational technique that allows any type of system to localize itself in an unknown environment and build the map of that environment, this process can be performed using GPS based localization, Visual SLAM [VSLAM] and LiDAR-SLAM. VSLAM is one of the widely used mechanism in autonomous system like robotics and vehicles to localize, build map from image captured by the

visual sensors for indoor scenarios [1]. The issue of localization and mapping becomes more complex because of the aspects like environmental changes, light condition variations and occlusions which have a strong influence on feature extraction and matching processes. Some of the widely used feature extractors in vSLAM include Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), Oriented FAST and Rotated BRIEF (ORB). These processes perform well in relatively minimal changing scenarios, issues arise when the environment is undergoing rapid change in real timeframe. Also, traditional methods fail to scale when the feature extraction process to be performed on large volumes of data and/or is high di-mensional and also when the features visual in nature are either sparse or under-go covariate shift.

The current work utilizes State-of-the-art deep learning approaches that can fa-cilitate feature extraction and matching in VSLAM.

The Detect and Describe Network deep learning architecture [D2-Net] is considered as a baseline model [3] upon which improvisations like the Squeeze and Excitation Blocks [SE-Blocks], Cross-Layer Feature Aggregation and Hierarchical Feature Progression have been implemented. The D2Net model performance by itself is not promising for certain vSLAM scenarios, like the high velocity movement condition, large scale environment, and restricted computational powers. This paper proposes a novel advanced D2Net architecture that supports feature extraction and matching improvements for visual odometry in scenarios that undergo rapid changes in real timeframe and those with adverse or poor lighting conditions. Model Perfor-mance evaluations are carried out on standard indoor dataset to test the effec-tiveness of the proposed enhanced D2-net architecture for Feature Extraction and mapping process where empirical results indicate that the obtained matching performance is better with reduced computational complexity in handling similar matches. This also formulates a foundation for next steps of the VSLAM pipeline including Pose and Depth Estimation, Local Mapping and Correction, Global mapping and optimisation and loop closure detection, so that VSLAM process can happen efficiently and effectively

II. LITERATURE SURVEY

The comprehensive overview of some of the related work has been discussed in this section. A comparative evaluation of ORB-SLAM3 and DynaSLAM was conducted on the KITTI and TUM RGB-D datasets for training, with monocular camera-captured datasets used for validation of the algorithms. The results indicate that while both algorithms perform effectively, DynaSLAM performs better than the ORB-SLAM3 in dynamic environments by detecting and excluding dynamic objects from SLAM processing and also observed that the VSLAM outperforms the LiDAR based SLAM for indoor scenarios. However, the study is limited to two algorithms and lacks detailed quantitative evaluations of computational complexity and resource consumption. These findings underscore the im-portance of addressing dynamic object handling and resource-efficient feature extraction, motivating the current proposed work to enhance robustness and adaptability in VSLAM systems [2]. In parallel, a trainable CNN for joint detection and description of local features, known as D2-Net, was introduced. This model combines detection and description tasks into a single pipeline, achieving significant improvements in robustness to scale and viewpoint changes. Despite its strong performance, the original D2-Net does not include optimizations for computational efficiency or specific adaptations for dynamic scenarios. These insights influenced the enhancements in the proposed work, including additional convolutional layers and dilated convolutions, to better handle complex environments and improve matching accuracy [3]. Furthermore, a detailed classification of state-of-the-art VSLAM methods that integrate deep learning for visual odometry, loop closure detection, and mapping was provided.

The survey also reviews evaluation metrics and datasets, demonstrating the benefits of deep learning for improved robustness and processing speed. However, the limitations of deep learning methods, such as generalization, real-time performance constraints, and a lack of standardized benchmarking frameworks, are identified. These challenges are directly addressed in the proposed work by optimizing the deep learning-based feature extraction pipeline for dynamic environments [4]. Additionally, semantic object annotations were incorporated into a visual SLAM framework using an enhanced YOLO-v3 for real-time object detection. The system achieved high processing speeds while maintaining SLAM functionality in both indoor and outdoor environments. However, its reliance on YOLO-v3 for object localization under challenging conditions, such as occlusions or poor lighting, limits its robustness. This limitation underscores the need for more robust feature extraction methods, as implemented in the proposed D2-Net enhancements, which include attention mechanisms for improved adaptability to complex scenarios [5]. In another development, OFM-SLAM, a visual semantic SLAM system based on ORB-SLAM2 and YOLOv3 for detecting and tracking dynamic objects, was developed. While the system demonstrated superior efficiency in tracking dynamic objects compared to ORB-SLAM2, it

struggled with error propagation and lacked validation in complex outdoor scenarios. These findings highlight the importance of integrating feature extraction and dynamic object handling into a unified framework, a goal addressed in the proposed work through advanced model design and optimization [6]. The Local Feature Transformer [LoFTR], a detector-free model utilizing transformers for local feature matching, was introduced. LoFTR replaces traditional feature detectors with a self-attention mechanism that directly matches dense features between image pairs. This architecture significantly enhances matching accuracy, especially in challenging environments involving complex textures or large viewpoint changes. However, the computational cost of transformers poses challenges for deployment in real-time or resource-constrained scenarios. These insights emphasize the need for computationally efficient models, motivating the enhancements in the proposed D2-Net-based pipeline to balance accuracy and performance in VSLAM systems [7].

Considering the reviewed literature, it can be noted that based on deep learning techniques several algorithms for feature extraction and matching in VSLAM exists, critically several gaps need to be addressed. Existing methods often fail to generalize effectively across diverse environments, particularly in handling dynamic and visually ambiguous scenarios [8] [9].

This limitation highlights the need for more robust features that are not only resilient but also interpretable for users with limited expertise in the field. Furthermore, deep learning models continue to face challenges in achieving real-time processing on resource-constrained platforms due to their computational complexity. Another issue is the limited transparency in how these models make decisions [11], which is a critical barrier to their adoption in practical applications, the lack of standardized benchmarks and universally accepted indices for quantitatively assessing feature extraction and matching approaches hampers the ability to compare advancements across the field comprehensively. Addressing these challenges through innovative deep learning-based methodologies, such as those focusing on enhanced feature extraction and dynamic object handling [10] [12] [13], is crucial for advancing VSLAM, enabling reliable visual odometry, for achieving precise localization and mapping in complex and dynamic settings.

III. METHODOLOGY

A. Details of the Dataset used [1]

- Description: TUM RGB-D is widely used dataset for testing visual SLAM systems in indoor scenarios, which contains RGB and depth images and ground truth for system trajectory.
- Sequence Used: fr1/xyz; 798 RGB-D images
- Content: The fr1/xyz sequence includes RGB and depth images [obtained from an RGB-D camera like Microsoft Kinect]. It provides number of indoor sequences, which are desirable for evaluating feature extraction and matching algorithms in heavily structured settings.
- Resolution: These images are captured at a screen resolution of 640 X 480.
- Ground Truth: The data set can supply vital information in the comparison of actual and obtaining trajectory, so visual odometry and SLAM technologies performance can be estimated successfully

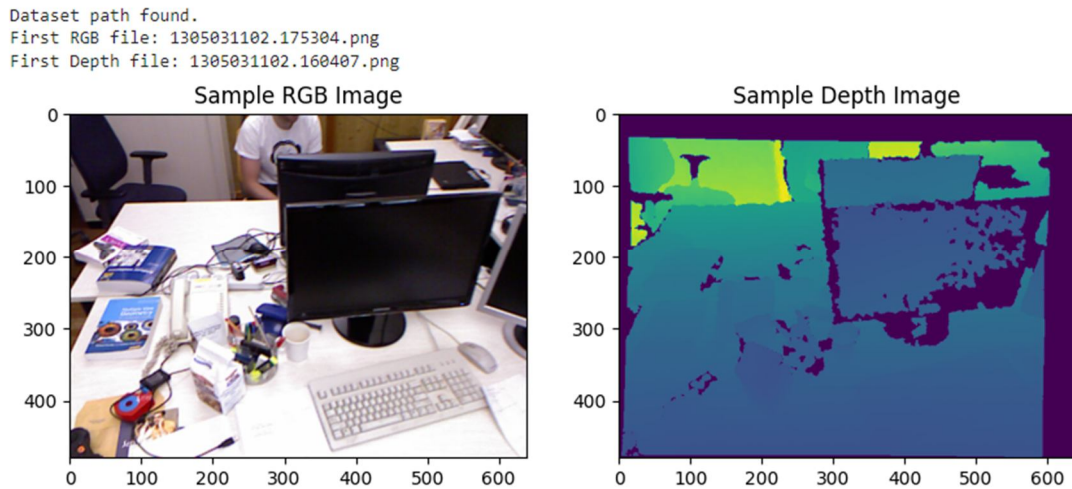


Figure 1. Dataset Loading and verification results while setting up the environment

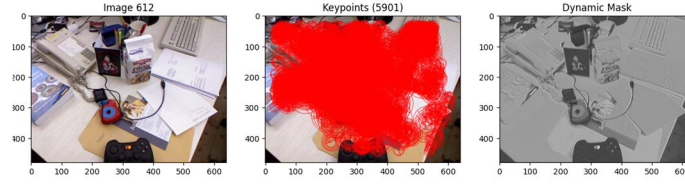


Figure 2 Illustration of Keypoints and Dynamic masks from the proposed algorithm

B. Details of Models Implemented

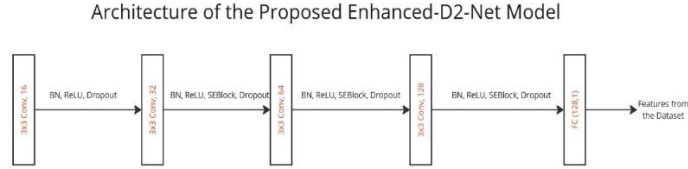


Figure 3 Architecture of the proposed Enhanced D2Net Model

The Figure 3 provides the insight to proposed model for feature extraction and feature matching in visual odometry uses downsized RGB images of dimension 640 X 480 from the dataset, these are convoluted and pooled to extract the features at multiple abstraction levels using a number of convolutional layers. The initial layer of convolution (Conv1) is comprised from 16 filters, which then goes through Batch Normalization and ReLU activation to learn some of the fundamental features of the images such as edges and Texture. The second convolutional layer involves convolving the input through 32 filters deepening the network ability to learn more complex features and thus elementarily we apply batch normalization followed by ReLU activation. The feature extraction process is advanced in the third convolutional layer, the Conv3, which uses 64 filters to parse the image for advanced features. The last Convolutional layer (Conv4) uses 128 feature maps and applying a dilation factor of two, which allows the network to grasp the global context characteristics of a scene from a large field of view which is useful in complicated scenes. To enhance the network ability to consider the important aspects, Squeeze-and-Excitation blocks are implemented after the convolutional layers. These blocks are composed of two layers of fully connected layers that are employed to recalibrate the obtained feature maps based on the inter-channel relations and as a result, augment the network capacity to attend for the most significant features and downplay the less important ones. After that, using the stream of feature maps, MaxPooling layer is added in order to diminish the size of the feature maps to prevent excessive computations while pre-serving important characteristics. There is a pooling layer followed by a Dropout layer to prevent overtraining with some of the input units set to zero as a training technique. The last layer of the network is the Fully Connected Layer [FC Layer] that maps the features learned into a lower dimensional vector. This feature vector is used as the descriptors or the key points of the input image, for feature matching used in visual odometry and VSLAM. Including these layers, the total forecasted trainable parameters of the proposed model are 3,943,944. This is due to the fact that the proposed method, being specifically designed for feature extraction and matching, can be implemented for real-time use in dynamic environments that are the characteristics used by VSLAM, where computational resources are usually limited.

Feature Matching Pipeline of Proposed Focus-Matcher

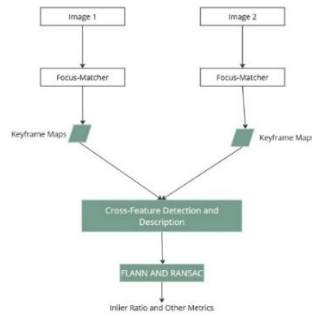


Figure 4. Focus Matching Pipeline of the proposed Focus-Matcher

IV. RESULTS

This section describes the evaluation of the proposed Enhanced D2Net model for the Freiburg1/xyz RGB-D dataset. The performance evaluation is made in terms of feature extraction time, feature maps and inlier ratio.

A. Feature Extraction and Feature Maps

The odometry of the Enhanced D2Net model uses convolutional layers to obtain features of the RGB images in a hierarchical methodology. Furthermore, attention modules namely Squeeze-and-Excitation (SE) blocks are integrated to improve the feature representation by recalibrating the feature importance.

The feature extraction was proved to be very efficient; the model analyzed all 798 images within 60.46 seconds, or 76 ms seconds per image achieving near real-time performance. Throughout the feature extraction phase, there was a hierarchical transformation of features concerning layers from simple texture descriptors to more complex semantically related features. At higher layers, the feature maps involved relatively larger scale regions, which again helped the system to operate effectively in complex scenarios involving changes in the environment. These coarsened feature maps play an important aspect of feature matching and are attributed largely to the performance of the model in a real world.

B. Inlier Ratio

The Inlier Ratio gives the ratio of inliers to the total number of feature matches for a given model after the process of RANSAC. The model demonstrated very high results in the inlier ratios equal to 97.72% for all consecutive image pairs, which means that the model was successfully determining the correct feature matches in every consecutive image pair configured with noise and dynamism factors also with the complete pipeline being executed within 63.6 ms, refer Table 1 shows various other metrics obtained. This result showcases the model's benefit for correctly matching features as most of them are classified as inliers, making feature correspondences accurate for tasks such as localization and Mapping.

C. Good Matches

In total, the model obtained successful alignments of 4,80,239 good feature matches for all consecutive image pairs with RANSAC inliers being at 4,69,271 and 602 good matches per image pairs, with the numbers of RANSAC inliers being at 589. The representation of good feature matches per image pair as shown in Figure 6 for the considered dataset. Such a high number of good matches proves that the model manages to extract necessary correctly mapped feature points, which are essential to provide stable visual SLAM.

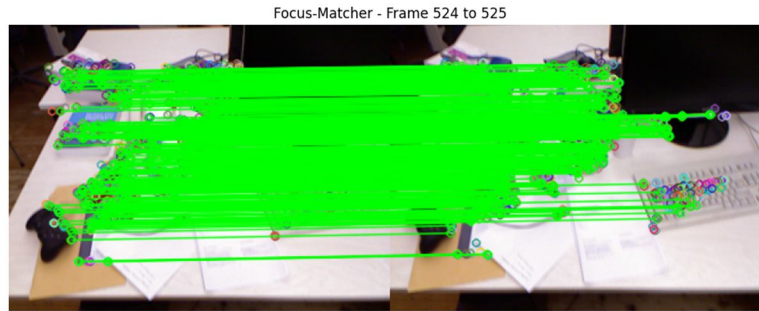


Figure 5 Feature Matching of the proposed Focus-Matcher between two consecutive frames

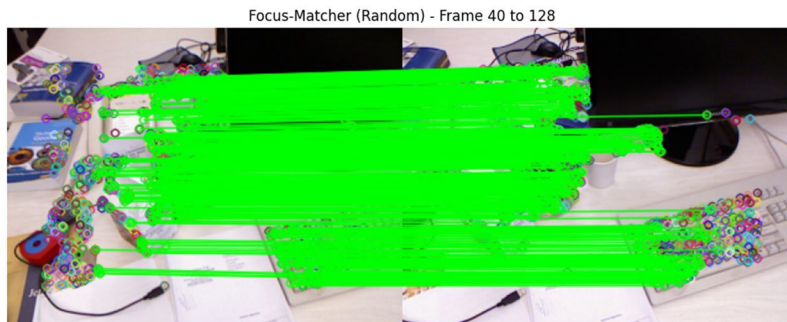


Figure 6 Feature Matching between two random frames of the proposed Focus Matcher

TABLE I. METRICS OBTAINED FROM IMPLEMENTATION OF THE PROPOSED FEATURE MATCHER ALGORITHM

Metric	Value
Total Pairs Processed	797
Average Keypoints Detected	1638
Cumulative Total Matches (Ratio Test)	4,80,239
Cumulative Total Matches (RANSAC Inliers)	4,69,271
Average Matches Per Pair (Ratio Test)	602
Average Matches Per Pair (RANSAC Inliers)	589
Average Outliers Rejected by RANSAC Per Pair	13
Average Inlier Ratio (RANSAC vs Ratio Test)	97.72%
Average Feature Detection Time	0.0168 sec
Average Feature Matching Time (KNN + Ratio)	0.0425 sec
Average RANSAC (Homography) Time	0.0021 sec
Average Total Processing Time Per Pair	0.0636 sec

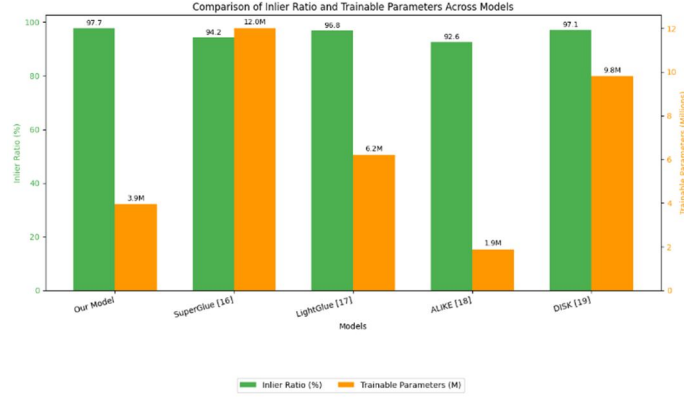


Figure 5 Comparison of Inlier Ratio and Trainable parameters across models

V. DISCUSSION

The efficiency of the feature extraction in the proposed model was established to be optimal as all the 798 RGB images were processed within the acceptable time frame required for localization. This efficiency coupled with high accuracy of the feature matching shows that the proposed model can easily work for larger data with relatively small computational overhead. When analyzing the feature map, we are also able to get more insight on how the network is successful in learning large number of features which are also hierarchical. The feature matching was slowly building up layers from the primary component of an image to higher im-age abstractions that is, the activations are spread widely throughout all the layers of the model and this is a sign of good generalization in part because good models learn a variety of features.

This order of the features presented in feature maps gives a perception that the model can develop a hierarchical concept over the input data. This hierarchical learning is crucial for other vision applications requiring both the local and the global context. Also, because the learned features are different in each layer, one can say that overfitting is minimized, and the target information for the further steps of the pipeline is properly extracted. Further, the proposed model generated an exceptionally high number of good matches, resulting in an inlier ratio of 97.76%. While the other models go very high in terms of number of parameters to be trained and computational resources the proposed model balances number of parameters and inlier ratio- with only 3.9M parameters the model exhibits better performance than the SOTA [State of the Art] models like SuperGlue[16], LightGlue[17] and DISK[19]. This is a clear sign of the model's capacity to identify reliable key point correspondences, which in turn raises confidence in feature matching approaches. The results of this work suggest that the proposed model is computationally efficient and is able to learn the necessary high-quality features for good matching.

VI. CONCLUSION

The proposed model effectively combines a balanced architecture with advanced feature extraction techniques, particularly tailored for RGB image datasets. By incorporating convolutional layers, batch normalization, and squeeze-and-excitation blocks, the model optimizes feature maps to build high-level, hierarchical features. The

process begins with a generic preprocessing stream for RGB images, allowing the model to extract features at various levels of granularity. Attention mechanisms applied at different stages further enhance the focus on significant attributes, improving the model's overall efficiency and robustness.

This approach demonstrates strong applicability in handling dynamic features across image sequences, making it ideal for use in Visual SLAM (Simultaneous Localization and Mapping) and other computer vision tasks. Beyond basic feature extraction, the model's adaptability also extends to mapping, loop closure detection, and localization in complex environments, contributing significantly to the development of more robust SLAM systems. This integration of deep learning-based feature extraction and attention mechanisms enables better performance in challenging real-time scenarios, particularly in dynamic and resource-constrained settings. Future work in this pipeline involves pose and depth estimation, local mapping, alignment and optimisation, loop closure, global map generation and semantic information extraction.

REFERENCES

- [1] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A Benchmark for the Evaluation of RGB-D SLAM Systems," in *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**, San Francisco, CA, USA, 2012, pp. 573–580.
- [2] K. J. de Jesus, M. O. K. Pereira, L. R. Emmendorfer, and D. F. T. Gamarra, "A Comparison of Visual SLAM Algorithms ORB-SLAM3 and DynaSLAM on KITTI and TUM Monocular Datasets," in *Proceedings of the 2023 SIBGRAPI Conference on Graphics, Patterns, and Images*, 2023, pp. 109–114. IEEE.
- [3] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-Net: A Trainable CNN for Joint Detection and Description of Local Features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8092–8101.
- [4] Y. Zhang, M. Liu, H. Li, and H. Li, "A Survey of Deep Learning Technology in Visual SLAM," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- [5] M. Xu, Q. Liu, X. Zhang, and X. Chen, "Incorporating Object Semantic Annotations in-to Visual SLAM Framework with YOLO-v3," in *Proceedings of Springer Robotics and Autonomous Systems*, vol. 148, pp. 24–37, 2022.
- [6] Z. Xiong, Y. Zhao, M. Q. Anwar, and T. Zhou, "OFM-SLAM: Dynamic Object Identification and Map Formation with ORB-SLAM2 and YOLOv3," *Journal of Visual Navigation*, vol. 11, no. 4, pp. 334–346, 2021.
- [7] Z. Zhang, Y. Sun, S. Luo, G. Sattler, and Y. Liu, "LoFTR: Detector-Free Local Feature Matching with Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8923–8932.
- [8] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A Benchmark for the Evaluation of RGB-D SLAM Systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, 2012, pp. 573–580, Oct.
- [9] S. Mokssit, D. B. Licea, B. Guermah, and M. Ghogho, "Deep Learning Techniques for Visual SLAM: A Survey," *IEEE Access*, vol. 11, pp. 20026–20050, 2023.
- [10] Y. Zhang, Y. Wu, K. Tong, H. Chen, and Y. Yuan, "Review of Visual Simultaneous Localization and Mapping Based on Deep Learning," *Remote Sensing*, vol. 15, no. 11, p. 2740, 2023.
- [11] J. Han, K. Tang, H. Li, and S. Jiang, "Indoor Positioning Systems of Mobile Robots: A Review," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 15340–15355, 2021.
- [12] Y. Yang, G. Zhang, and H. Bao, "Deep Learning for Visual Localization and Mapping: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3365–3387, 2022.
- [13] J. A. Bala, S. A. Adeshina, and A. M. Aibinu, "Advances in Visual SLAM Techniques for Autonomous Vehicles: A Review," *IEEE Access*, vol. 9, pp. 107710–107737, 2021.
- [14] D. Gamarra, M. Kesse, L. Marin, and T. Oliveira, "Comparing ORB-SLAM3 and DynaSLAM Using Monocular Vision," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2021, pp. 1234–1240.
- [15] X. Zhao, T. Zuo, and X. Hu, "OFM-SLAM: A visual semantic SLAM for dynamic in-door environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021. DOI: 10.1155/2021/5538840.
- [16] 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 (CVPR)*, 2020, pp. 4938–4947. DOI: 10.1109/CVPR42600.2020.004938.
- [17] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 17627–17638. DOI: 10.1109/ICCV51965.2023.017627.
- [18] X. Zhao, Y. Wu, J. Miao, W. Chen, P.C.Y. Chen, and Z.G. Li, "ALIKE: Accurate and lightweight keypoint detection and descriptor extraction," *IEEE Transactions on Multi-media*, vol. 25, pp. 3101–3112, 2023.
- [19] M.J. Tyszkiewicz, P. Fua, and E. Trulls, "DISK: Learning local features with policy gradient," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 14254–14265, 2020.