MIAS: An Efficient Active Multi-Agent Distributed Training System for Robotic Implicit SLAM

Your N. Here
Your Institution

Second Name
Second Institution

Abstract

Online training tasks rely on training input consecutively sampled from the real world to refine their training model, and automated mobile devices (robots) shouldering the responsibility of sampling has the potential to further enhance their training performance by actively sampling in reaction to the real-time training quality, besides saving human labor. However, to achieve such active sampling faces two major gaps. First, as the training quality is typically computation-intensive to get for a robot (e.g., validating), where to move for the robot to boost training performance? Second, since the consecutive samples from nearby states (e.g., position, orientation) of a robot are too similar and would limit the training information gain, how to move for better efficiency?

We observe that real-time training loss is spatially and temporally related to locations in the environment and implies the level of information gain on further sampling of these locations. The second gap can be overcome by cooperative multi-view sampling from multiple robots, since they are naturally distant in state space. Based on these observation, we choose to build an online training application named MIAS (Multi-robot Implicit Active SLAM) that drives multiple robots to actively and cooperatively sample the environment in real-time reaction to the quality of the training implicit SLAM model. Evaluation shows that MIAS with three robots speeds up the implicit SLAM tasks not only by up to xxX compared to the baselines with three robots, but also by xxX compared to MIAS with single robot.

1 Introduction

Online training tasks (e.g., implicit SLAM, domain adaptation, long term learning) take unlabelled training input consecutive sampled from the real world to refine their training model for better performance in the changing real world environments. The sampling of such training input typically relies on human labor (e.g., the RGBD image sequences captured by hand-held camera in implicit SLAM); offloading such online training tasks to automated mobile devices (robots equipped with GPUs) not only enables automated sampling, which saves human labor, but also has the potential to enable active sampling in reaction to the real-time training quality to boost training performance (e.g., sample the training input with highest training information gain).

However, enabling active sampling for online training tasks deployed over robots faces two major gaps. First, where to move for higher training information gain? Typically, quantifying the training quality of the training model relies on statistical methods such as validating, which is computation-intensive for robots and infects the ongoing online training task. Without the real-time statistics of training quality, it is difficult to decide the destination for the robot to mobile in active sampling.

Second, how to move for higher training information gain? The consecutive sampling of a robot suffers the problem of sampling locality: while the robot is moving in local state space (i.e., nearby position and orientation), the consecutive samples are often too similar, limiting training information gain. We visualize in Fig.**??** the training information gain of consecutive training input about an area of interest (e.g., a wall that has high potential training information gain) after a robot sampling this area once in an implicit SLAM task and we can find that there is a lowland of training information gain around the starting state of the robot. As a result, to complete sampling of an area of interest, the robot has to move in and out the lowland of training information gain during sampling, lowering the average training information gain per sample.

To overcome to above gaps, we observe that real-time training loss is spatially and temporally related to locations in the environment and implies the level of information gain on further sampling of these locations. ALthough training loss does not directly reveal the statistics of training model quality such as accuracy, but it is semantically related with state space when the robot sampled the corresponding training input. A higher training loss compared with others reveals the training model is more inaccurate against the corresponding training input and higher possible level of training information gain of further sampling in the corresponding state space, which has the potential to guide the navigation for active sampling.

For the second gap, we find that multiple robots are naturally distant in state space and their sampling from multiple perspectives (multi-view sampling) has the potential the skip the lowland of training information and mitigate the locality problem. With multi-view sampling, it is possible that after one robot samples an area of interest, other distant robots would shoulder the responsibility of further sampling the same area (if loss is still high) beyond the lowland of training information gain and the first robot could quickly switch its target and avoid the lowland.

With the above ideas, we take the first step to build a multi-robot active online training application and choose implicit SLAM as the main workload, naming MIAS, Multi-view Implicit Active SLAM. The implicit SLAM builds dense mesh of the surrounding environment in real time by optimizing both the localization of the state (position and orientation) of the robot (i.e., tracking) and an implicit representation of the dense mesh (i.e., mapping) over consecutive RGBD images sampled from robot cameras. Traditional methods to automate the sampling for SLAM are unfit for implicit SLAM: first, they trace the explicit representation of the map (e.g., the dense mesh) to estimate the quality of tracking and mapping for decision making, which could only be obtained by computation intensive validation of the implicit representation in implicit SLAM, slowing down the decision making; second, they typically avoid multi-view sampling of multiple robots so that the robots would sample different areas to optimize coverage, suffering the sampling locality problem.

But with MIAS, we can achieve decision making in real-time reaction to online training quality by tracing the change of loss level: when the training loss for tracking increases, the state of the robot is inaccurate and we control the robot to sample the previously sampled area for re-localization; when the training loss for mapping increases, we mark the corresponding areas as places of interest for further sampling; when both losses are low, we explore new areas. The sampling of places of interest is further accelerated by multi-view sampling beyond the lowland of information gain. As a result, both training quality and training information gain per sample are optimized.

We implemented MIAS on four-wheel robots equipped with RGBD cameras and state-of-the-art (SOTA) mobile GPU chips. The implicit representation of dense mesh of implicit SLAM is distributedly trained among the robots over SOTA distributed training library optimized for mobile devices with gradient compression. We find that the cost for distributed training (e.g., time for compressing and communicating the gradients) is little for a single robot, accounting for only xx% time of the computation of gradients. We compare MIAS with two traditional active SLAM methods and evaluated over a team of three robots in both habitat-simulated environments and real-world environments. Evaluation shows that:

- MIAS accelerates the active implicit SLAM process. With the same time cost, MIAS increased the accuracy of the built dense mesh by xxX to xxX compared with the baselines. When reaching the same high accuracy of the built dense mesh, MIAS reduced the total time cost by xx% to xx%.

- N robots involved brings N+ times acceleration. When reaching the same high accuracy of the built dense mesh, MIAS with two / three robots reduced the total time cost by xx% to xx% / xx% to xx% compared with the baselines and MIAS with one robot. With the same time cost, MIAS with two / three robots increased the accuracy of the built dense mesh by xxX to xxX / xxX to xxX.

Our major contributions are the paradigm of enabling active sampling for online training tasks deployed over robots and the real-world multi-robot active implicit SLAM application, MIAS. The paradigm would benefit various online training tasks deployed on robots such as long term training or domain adaptation in the filed by bridging the real-time interaction between online training and robot mobility, so that the mobility of robots can better serve for optimizing the training quality of online training tasks via action decision making based on real-time training loss and optimizing the training information gain per sample via multi-view sampling. The application MIAS not only automatically samples the environment to build dense mesh of the environment in real time that saves human labor, but also optimizes the quality of dense mesh and the training information gain per sample by active sampling, leading to acceleration of dense mesh building beyond the number of robots involved.

The rest of this paper is organized as follows:...

2 Background & RELATED WORK

2.1 Online Training on Multi Robot

Machine learning (ML) approaches are generally trained for a specific task on a dedicated training set. However, in many real-world applications, Labeling datasets are very expensive, and the data distributions can differ or even change over time. Therefore, Some unsupervised methods are proposed to learn knowledge from unlabeled data and make the machine learning model to adapt the new dynamic environment. For example, dynamic unsupervised domain adaptation methods [25] is proposed to adapt a pretrained model to a new environment by training it with both unlabeled data from the dynamic environment.

With the rapid development of such methods, robots can adapt their pretrained models to new scenarios(e.g., domain shifts or changing data distributions) after training with online collected data to retain the high accuracy of the models. As another example, neural implicit representations have recently become popular in simultaneous localization and mapping (SLAM), especially in dense visual SLAM. This method enables high-fidelity and dense 3D scene reconstruction by collecting unlabeled image sequences with RGB-D sensors in real-time. We envision the prosperity of these multi-robot collaborations and unsupervised learning methods are making online training on real-time collected data on multi-robot realistic.

2.2 Multi-robot System

For time-sensitive Search and Rescue (SaR) missions, using multiple cooperative robots is useful since they allow for quick environment exploration and offer more redundancy than using a single robot. A fully distributed SLAM system for robotic teams that can identify inter-robot loop closures without exchanging raw data was proposed by Pierre-Yves Lajoie et al [8].A completely distributed multi-robot system for dense metric-semantic SLAM is proposed by Yun Chang et al. [2] Every robot creates a local mesh and a local trajectory estimate. When two robots are in communication range, they start a distributed place identification and resilient pose graph optimization process.A multi-robot system is proposed to complete the online training task, each robot is equipped with a Jetson Xvaier NX for computation.

2.3 Related Work

Dense Visual SLAM. Visual SLAM is an online approach that incrementally creates the map of an environment while localizing the robot within it. Meanwhile, it is an area that has received much attention in both industry and academia. Specifically, sparse visual SLAM algorithms estimate accurate camera poses and only have sparse point clouds as the map representation, While sparse visual SLAM algorithms estimate accurate camera poses and only have sparse point clouds as the map representation, dense visual SLAM approaches focus on recovering a dense map of a scene, which makes the method very suitable for 3D reconstruction. Dense tracking and mapping (DTAM), proposed by Newcombe et al. [15], was the first fully direct method in the literature.

Neural Implicit-based SLAM. Neural implicit representations [13] have shown great performance in many different tasks, including 3D reconstruction [10, 12, 16, 17], scene completion [7, 9, 18], novel view synthesis [11, 19, 20, 24, 26], etc. In terms of SLAM-related applications, some works [3, 5] try to jointly optimize a neural radiance field and camera poses, but they are not suitable for large objects or wide range of camera motion. In addition, some recent works [4, 21] can support large-scale mapping, but they mainly rely on state-of-the-art SLAM systems like ORB-SLAM to obtain accurate camera poses, and do not produce 3D dense reconstruction.

NICE-SLAM [28] and iMAP [23] are the most famous two SLAM pipelines using neural implicit representations for both mapping and camera tracking. Since iMAP uses a single MLP as the scene representation so they are only adapt to small scenes,whereas NICE-SLAM, which uses hierarchical feature grids and small MLPs as the scene representation, can scale up to considerably bigger interior spaces. Nevertheless, it calls for RGB-D inputs, which restricts their use in outdoor settings or when only RGB sensors are available. In order to solve this problem, a new work named NICER-SLAM [27]was proposed, which is the first dense RGB-only SLAM, optimizes mapping and tracking end-to-end and also allows the high-quality synthesis of new views.

Active Mapping/SLAM. In the interest of exploring the environment by planning the path of mobile robots, active SLAM combines SLAM with path planning. This improves and speeds up the SLAM algorithm's ability to produce high-precision maps. The three active vision issues (localization, mapping, and planning) are combined by active SLAM. Robots can now autonomously carry out localization and mapping tasks, which helps to improve the accuracy of both those tasks and the representation of the environment. This topic has been studied before [6] came up with the phrase "Active SLAM," mostly as known as "exploration problems" [14, 22].

Specifically, iRotate [1] offers an active visual SLAM approach for omnidirectional robots because the static camera restricts the freedom of visual information acquisition. During the path execution, the robot can actively and continuously control its camera heading to maximize the environment coverage by taking advantage of its omnidirectional nature. The robot can significantly speed up the information-gathering process and quickly reduce the level of map uncertainty by actively performing coverage. In particular, these methods need to explicitly build maps before they can work, so they cannot be directly applied to the implicit SLAM framework. At the same time, the memory overhead of building explicit maps is large, and the lack of memory resources of robots often cannot support such active SLAM methods.

3 Overview

This chapter presents the architecture of MIAS and gives an overview of how MIAS achieves real-decision making in reaction to real-time training quality and cooperative multi-view sampling across robots.

References

[1] Elia Bonetto, Pascal Goldschmid, Michael Pabst, Michael J Black, and Aamir Ahmad. irotate: Active visual slam for omnidirectional robots. *Robotics and Autonomous Systems*, 154:104102, 2022.

[2] Yun Chang, Yulun Tian, Jonathan P. How, and Luca Carlone. Kimera-multi: a system for distributed multi-robot metric-semantic simultaneous localization and mapping.

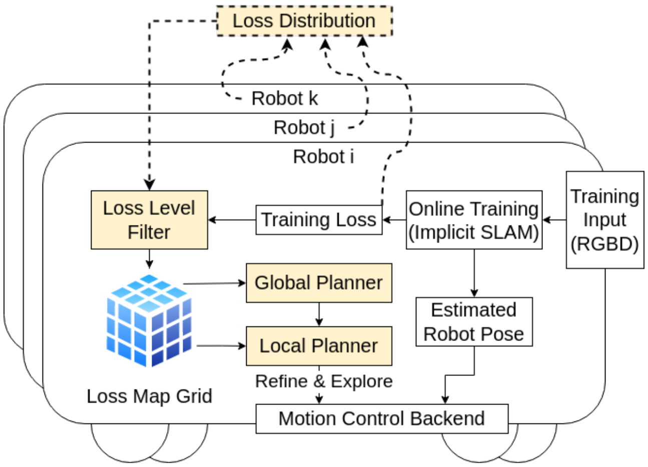

Figure 1: Overview of MIAS

In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11210–11218, 2021.

[3] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 264–280. Springer, 2022.

[4] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang-Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H Hsu. Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. *arXiv preprint arXiv:2209.13274*, 2022.

[5] Ronald Clark. Volumetric bundle adjustment for online photorealistic scene capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6124–6132, 2022.

[6] A.J. Davison and D.W. Murray. Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):865–880, 2002.

[7] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6001–6010, 2020.

[8] Pierre-Yves Lajoie, Benjamin Ramtoula, Yun Chang, Luca Carlone, and Giovanni Beltrame. Door-slam: Distributed, online, and outlier resilient slam for robotic teams. *IEEE Robotics and Automation Letters*, 5(2):1656–1663, 2020.

[9] Stefan Lionar, Daniil Emtsev, Dusan Svilarkovic, and Songyou Peng. Dynamic plane convolutional occupancy networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1829–1838, 2021.

[10] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020.

[11] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[12] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

[13] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[14] JE Moody, SJ Hanson, and RP Lippmann. Active exploration in dynamic environments. In *Advances in Neural Information Processing Systems 4*. Citeseer, 1992.

[15] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011.

[16] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[17] Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. *Advances in Neural Information Processing Systems*, 34:13032–13044, 2021.

[18] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional

occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020.

[19] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[20] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.

[21] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022.

[22] Cyrill Stachniss, Dirk Hahnel, and Wolfram Burgard. Exploration with active loop-closing for fastslam. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 2, pages 1505–1510. IEEE, 2004.

[23] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021.

[24] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

[25] Qing Tian, Yanan Zhu, Heyang Sun, Songcan Chen, and Hujun Yin. Unsupervised domain adaptation through dynamically aligning both the feature and label spaces. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8562–8573, 2022.

[26] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.

[27] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. *arXiv preprint arXiv:2302.03594*, 2023.

[28] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022.