2025 IEEE International Conference on Robotics and Automation (ICRA)
May 19-23, 2025. Atlanta, USA

Uncertainty-Aware Visual-Inertial SLAM with Volumetric Occupancy Mapping

Jaehyung Jung[1], Simon Boche[1], Sebastián Barbas Laina[1], and Stefan Leutenegger[1]

***Abstract*— We propose visual-inertial simultaneous localization and mapping that tightly couples sparse reprojection errors, inertial measurement unit pre-integrals, and relative pose factors with dense volumetric occupancy mapping. Hereby depth predictions from a deep neural network are fused in a fully probabilistic manner. Specifically, our method is rigorously *uncertainty-aware*: first, we use depth and uncertainty predictions from a deep network not only from the robot's stereo rig, but we further probabilistically fuse motion stereo that provides depth information across a range of baselines, therefore drastically increasing mapping accuracy. Next, predicted and fused depth uncertainty propagates not only into occupancy probabilities but also into alignment factors between generated dense submaps that enter the probabilistic nonlinear least squares estimator. This submap representation offers globally consistent geometry at scale. Our method is thoroughly evaluated in two benchmark datasets, resulting in localization and mapping accuracy that exceeds the state of the art, while simultaneously offering volumetric occupancy directly usable for downstream robotic planning and control in real-time.**

I. INTRODUCTION

State estimation is a fundamental building block for downstream tasks including robot perception, human-robot interaction, and robot exploration. Cameras and inertial measurement units (IMUs) have become a typical minimalist sensor suite to approach state estimation problems by virtue of their complementary characteristics and ability to observe a metric scale. Visual-inertial simultaneous localization and mapping (VI-SLAM) is considered to be well-established with the recent advancement such as OKVIS 2 [1], ORB-SLAM3 [2], OpenVINS [3], and VINS-Fusion [4], to name a few. Reprojection or photometric errors formulated from a set of sparse or (semi-)dense landmarks are standard residuals in the VI-SLAM problem. However, this map representation is not suitable for higher-level robot tasks where densely populated free/occupied spaces should be modeled explicitly.

Dense mapping incrementally integrates onboard range or depth measurements given a sensor pose to a global model to build a volumetric map, which can use occupancy [5] or truncated signed distance field (TSDF) [6], [7] as the underlying map representation. Given that the observation model plays an important role in the reconstruction quality, one has to decide how much the measured range of depth contributes to each voxel based on the uncertainty. Ideally, the uncertainty should be statistically consistent. Moreover,

This work was supported by the European Commission Horizon Europe project AUTOASSESS (101120732) and DIGIFOREST (101070405).

[1]Smart Robotics Lab, School of Computation, Information and Technology (CIT); as well as Munich Institute of Robotics and Machine Intelligence (MIRMI) and Munich Center of Machine Learning (MCML), Technical University of Munich, Germany. firstname.surname@tum.de

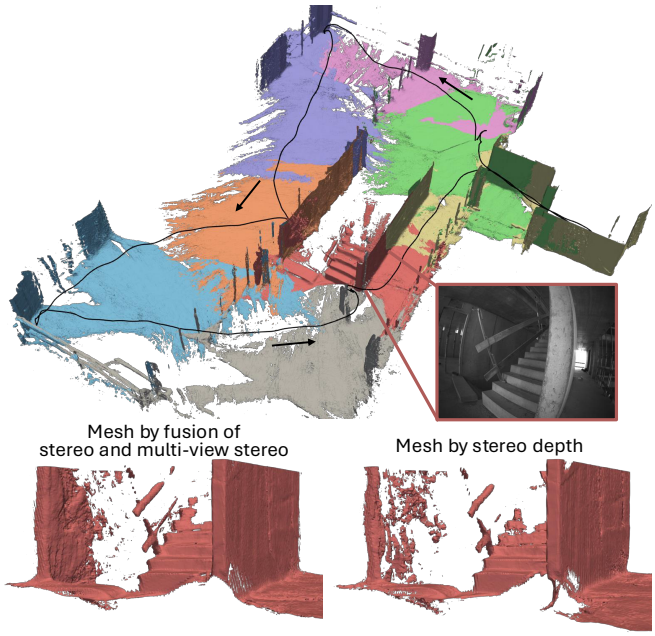

Fig. 1. Our method estimates pose (black line) and volumetric occupancy represented in submaps (visualised here as colored meshes). We fuse stereo and MVS network depths based on their predicted uncertainties giving less noisy meshes, as can be qualitatively observed in the left mesh. Result from the Hilti-Oxford dataset [10].

depth prediction with valid uncertainty or confidence has been an active research topic in computer vision [8], [9]. Depending on the camera configuration, a monocular depth network estimates depths in a single image, a stereo network predicts disparity between left and right images, and a multi-view stereo (MVS) network estimates depths in a reference image frame given multiple posed images. However, in contrast to the recent developments of uncertainty learning in deep neural networks, a quadratic model where the uncertainty increases with the squared depth has been widely used in dense mapping. This model, which effectively assumes constant disparity uncertainty, is a good approximation but is not valid in general. For instance, the belief in a highly uncertain pixel, as typically the case in less-textured areas, should be discounted for safe robot perception.

To tackle this challenge, we propose VI-SLAM with volumetric occupancy mapping which is fully *uncertainty-aware* in the sense that occupancy integration and the visual-inertial factor graph are properly weighted by the estimated uncertainty of *depth fusion*. We optimally fuse depths from stereo and MVS networks based on their predicted uncer-

2025 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3315-4139-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICRA55743.2025.11127824

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.

tainties motivated by the complementary characteristics of small and large baselines from static and motion stereo. As shown in Fig. 1, our approach outputs real-time robot pose as well as volumetric occupancy maps that are being built incrementally where submaps are anchored at a local coordinate frame for globally consistent map representation. In summary, our contributions of this paper are as follows.

- We introduce depth fusion of a deep MVS network and a stereo network based on their predicted network uncertainty to further improve the mapping accuracy.
- Leveraging the depth prediction and the associated uncertainty of a deep neural network, we formulate fully probabilistic VI-SLAM with volumetric occupancy submapping where the depth integration and the factor graph optimization are weighted by the predicted uncertainty.
- We thoroughly evaluate our approach in the benchmark datasets in terms of localization and mapping accuracy while achieving state-of-the-art performance running in real-time with a stereo-inertial setup.

II. RELATED WORK

We review the most relevant works in multi-view depth deep neural networks and dense SLAM including probabilistic mapping methods.

A stereo network aims to find the disparity in a rectified stereo image. Inspired by stereo matching [11], the network typically follows three steps: features are extracted from the stereo pair, a cost volume is built along the horizontal direction, and then the disparity is computed with the following refinement or upsampling. ACVNet [12] proposed attention concatenation volume to emphasize the most useful values in the cost volume. Unimatch [13] presented a unified network for stereo matching, optical flow, and two-view depth prediction with feature enhancement by a transformer [14]. Likewise, a deep MVS network follows a similar setup but with perspective warping based on camera geometry to construct the cost volume with unrectified images. MVSNet [15] aggregated multi-view features through a variance metric for the cost volume followed by 3D convolutions. Simplerecon [16] circumvented 3D convolutions by incorporating camera geometry in the cost volume. Simplemapping [17] proposed to input a prior depth map from sparse landmarks to narrow down the depth plane. Leveraging the recent developments in deep neural networks for multi-view depth prediction, we tailor state-of-the-art methods [13], [17] to learn the depth uncertainty via the Laplacian loss. The predicted uncertainty is utilized in our fully uncertainty-aware framework for depth fusion, occupancy mapping, and factor graph optimization.

Understanding dense geometry is indispensable for autonomous robots. KinectFusion [6] pioneered this area with TSDF fusion and frame-to-model registration. ElasticFusion [18] models an environment with dense surfels, which are updated non-rigidly after a loop. CNN-SLAM [19] and DeepFusion [20] employed a dense depth from a convolutional neural network and fuse depths from motion stereo

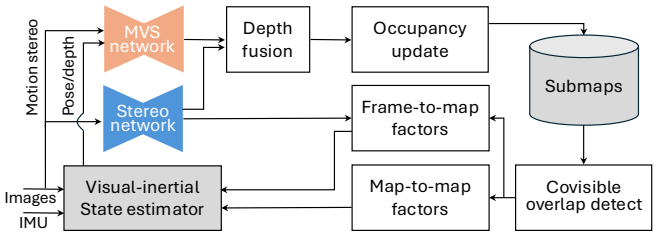

Fig. 2. An overview block diagram of the proposed method. We fuse depths from stereo and MVS networks from which the current submap is expanded. Fused depth and stereo depth formulate map-to-map and frame-to-map factors in the visual-inertial estimator.

to further improve the depth quality. Voxgraph [21] advocates a submap representation for a scalable system and proposed submap alignment based on Euclidean distance fields given odometry estimation. Regarding relevant works in an MVS network in SLAM, TANDEM [22] integrated a deep MVS network in visual odometry, rendering additional depths from TSDF for image alignment. Simplemapping adopted Simplerecon [16] and fused depths from the MVS network into a TSDF given poses and sparse landmarks from ORB-SLAM3 [2] but without any feedback from mapping. SigmaFusion [23] predicted dense depth uncertainty based on the information matrix from DroidSLAM [24] and showed less noisy reconstruction through uncertainty integration. However, their method is memory intensive and the mapping is decoupled from the estimator as an ad-hoc mapping as opposed to ours. D3VO [25] proposed to use the photometric uncertainty learned from depth estimation in direct visual odometry, but their method was not a full SLAM system. The most related work is [26] where an occupancy-to-point factor was introduced to align submaps. However, noisy depth measurements should be properly weighted for downstream tasks in contrast to precise LiDAR point clouds.

In contrast to previous works, our method tightly couples the visual-inertial estimator and volumetric occupancy with the occupancy-to-point factors for a globally consistent map. On top of this, our key contribution lies in the usage of uncertainty which is propagated to all downstream tasks.

III. DEFINITIONS AND NOTATIONS

The objective of this work is to estimate robot poses as well as volumetric occupancy values represented in a set of submaps given inertial and multiple camera measurements based on their observation uncertainty. A *submap* is defined as a local map representation with geometric overlaps to each other to reduce mapping error due to pose drift, since each submap's pose can be corrected after an update from the estimator [21], [27]. Our method is based on OKVIS2 [1] for visual-inertial estimation and Supereight2 [5] for occupancy submapping.

The world frame $\underrightarrow{\mathcal{F}}_W$ describes a gravity-aligned coordinate frame, the sensor frame $\underrightarrow{\mathcal{F}}_S$ is coincident with an IMU frame, and the ith camera frame $\underrightarrow{\mathcal{F}}_{C_i}$ is centered at the optical center of the corresponding camera. The jth submap frame $\underrightarrow{\mathcal{F}}_{M_j}$ is defined as the sensor frame when the new

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.

submap is declared. The robot state is represented by

$$\mathbf{x} = \begin{bmatrix} {}_W\mathbf{r}_S^T & \mathbf{q}_{WS}{}^T & {}_W\mathbf{v}^T & \mathbf{b}_\mathrm{g}^T & \mathbf{b}_\mathrm{a}^T \end{bmatrix}^T, \tag{1}$$

where ${}_W\mathbf{r}_S$ is the position of $\underrightarrow{\mathcal{F}}_S$ in $\underrightarrow{\mathcal{F}}_W$, $\mathbf{q}_{WS}$ is the Hamiltonian quaternion for orientation of $\underrightarrow{\mathcal{F}}_S$ relative to $\underrightarrow{\mathcal{F}}_W$, ${}_W\mathbf{v}$ is the velocity of $\underrightarrow{\mathcal{F}}_S$ relative to $\underrightarrow{\mathcal{F}}_W$, and $\mathbf{b}_\mathrm{g}$, $\mathbf{b}_\mathrm{a}$ are gyroscope and accelerometer biases, respectively. Using homogeneous representation (in *italic* font), the coordinate transformation from $\underrightarrow{\mathcal{F}}_B$ to $\underrightarrow{\mathcal{F}}_A$ can be written as ${}_A\boldsymbol{r}_S = \boldsymbol{T}_{AB}\,{}_B\boldsymbol{r}_S$ with $\boldsymbol{T}_{AB} \in \mathrm{SE}(3)$.

IV. VISUAL-INERTIAL ESTIMATOR

A. Overview

We introduce stereo and MVS networks, which predict depths as well as depth uncertainty, into visual-inertial state estimation to probabilistically integrate depths in occupancy mapping and weight occupancy-to-point (map-to-map and frame-to-map) factors in the factor graph optimization. Fig. 2 illustrates our approach: given multiple images and inertial measurements, we optimally fuse depths based on their predicted uncertainties. The fused depth map with its pose is integrated into the currently expanding submap. The most overlapping submaps to the latest submap are found based on its covisible submaps. Map-to-map factors are built from a pair of submaps, while live depths from the stereo network are used to compute the frame-to-map factor. These factors are weighted by predicted uncertainties from the networks. We describe each block in detail in the following sections.

B. Occupancy submapping

Each submap maintains occupancy log-odds in its coordinate frame where our submapping is based on Supereight2 [5], an octree-based adaptive multi-resolution mapping pipeline. Given the camera pose $\boldsymbol{T}_{MC}$ and the depth image $\mathbf{D}$, the log-odd of a point in the map frame, ${}_M\mathbf{p}$ is defined as

$$l\left({}_M\mathbf{p}\right) = \log\frac{P_\mathrm{occ}\left({}_M\mathbf{p} \mid \boldsymbol{T}_{MC}, \mathbf{D}\right)}{1 - P_\mathrm{occ}\left({}_M\mathbf{p} \mid \boldsymbol{T}_{MC}, \mathbf{D}\right)}, \tag{2}$$

where P_occ is the occupancy probability. To account for the uncertainty of depth measurements, we use the piece-wise linear inverse sensor model to approximate the log-odd

$$l\left({}_M\mathbf{p}\right) = \begin{cases} l_\mathrm{min} & d_r < -3\,\sigma_d, \\ \frac{|l_\mathrm{min}|}{3\sigma_d} d_r & -3\,\sigma_d \le d_r < \frac{\tau}{2}, \\ l_\mathrm{max} & \frac{\tau}{2} \le d_r < \tau. \end{cases} \tag{3}$$

Here, d_r stands for the distance between a queried point and the surface measurement along the ray, l_min is a configurable parameter for the minimum log-odd, σ_d is the standard deviation of the depth, τ stands for surface thickness, which is assumed to be proportional to the depth measurement, and l_max can be determined from the continuity. To integrate multiple measurements, the mean of log-odd $L(\cdot)$ is recursively updated as

$$\begin{aligned} L_k\left({}_M\mathbf{p}\right) &= \frac{L_{k-1}\left({}_M\mathbf{p}\right) w_{k-1} + l\left({}_M\mathbf{p}\right)}{w_{k-1} + 1}, \\ w_k &= \min\{w_{k-1} + 1,\ w_\mathrm{max}\}, \end{aligned} \tag{4}$$

where w_k is the number of observations and saturated in w_max. Supereight2 uses a heuristic uncertainty model to determine σ_d, such as a linear or a quadratic model. However, this approximation is not valid in general: for instance, textureless walls or object edges could cause highly uncertain depth prediction. In response, we propose to exploit pixel-wise learned uncertainty as an additional output of our deep neural network to integrate noisy depths into occupancy.

C. Factor graph optimization

Our visual-inertial state estimator is based on OKVIS2 [1], which is a visual-inertial SLAM formulated in a factor graph, and the extended work with LiDAR [26]. The estimator jointly minimizes the following cost function

$$\begin{aligned} c(\mathbf{x}) = &\frac{1}{2}\sum_i \sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{J}(i,k)} \rho_\mathrm{c}\left(\mathbf{e}_\mathrm{r}^{i,j,k\,T}\mathbf{W}_\mathrm{r}\mathbf{e}_\mathrm{r}^{i,j,k}\right) \\ &+ \frac{1}{2}\sum_{k\in\mathcal{P}\cup\mathcal{K}\setminus f} \mathbf{e}_\mathrm{s}^{k\,T}\mathbf{W}_\mathrm{s}^k\mathbf{e}_\mathrm{s}^k + \frac{1}{2}\sum_{r\in\mathcal{P}}\sum_{c\in\mathcal{C}(r)} \mathbf{e}_\mathrm{p}^{r,c\,T}\mathbf{W}_\mathrm{p}^{r,c}\mathbf{e}_\mathrm{p}^{r,c} \\ &+ \frac{1}{2}\sum_{k\in\mathcal{P}\cup\mathcal{K}} \rho_\mathrm{t}\left(\mathbf{e}_\mathrm{o2p}^{k,o(k)\,T}\mathbf{W}_\mathrm{o2p}^{k,o(k)}\mathbf{e}_\mathrm{o2p}^{k,o(k)}\right) \\ &+ \frac{1}{2}\sum_{m\in\mathcal{M}} \rho_\mathrm{t}\left(\mathbf{e}_\mathrm{o2p}^{m,o(m)\,T}\mathbf{W}_\mathrm{o2p}^{m,o(m)}\mathbf{e}_\mathrm{o2p}^{m,o(m)}\right). \end{aligned} \tag{5}$$

In this expression, the reprojection error $\mathbf{e}_\mathrm{r}^{i,j,k}$ with the Cauchy robustifier $\rho_\mathrm{c}(\cdot)$ runs over the ith camera, the set $\mathcal{K}$ including recent frames and keyframes in the past, and the visible landmark set in those frames $\mathcal{J}(i,k)$. The IMU preintegration error $\mathbf{e}_\mathrm{s}^k$ is formulated starting from poses of $\mathcal{K}$ and the pose graph set $\mathcal{P}$ except from the current frame f. The relative pose error $\mathbf{e}_\mathrm{p}^{r,c}$ between the rth pose graph frame and the cth frame in the set $\mathcal{C}(r)$, which has edges to r, is also added. The weighting matrices $\mathbf{W}_\mathrm{r}, \mathbf{W}_\mathrm{s}, \mathbf{W}_\mathrm{p}$ are obtained from the inverse of the reprojection covariance matrix, IMU uncertainty propagation, and the landmark marginalization, respectively. The occupancy-to-point factor $\mathbf{e}_\mathrm{o2p}$ with the Tukey loss function $\rho_\mathrm{t}(\cdot)$ encodes the fact that the point should lie on the surface ($L = 0$), and it includes N depth measurements as

$$\begin{aligned} \mathbf{e}_\mathrm{o2p}^{a,b} &= \begin{bmatrix} e_{\mathrm{o2p}_1} & \cdots & e_{\mathrm{o2p}_N} \end{bmatrix}^T, \\ e_\mathrm{o2p} &= \frac{L\left(\boldsymbol{T}_{S_b W}\boldsymbol{T}_{W S_a}\,{}_{S_a}\boldsymbol{p}\right)}{\left\|\nabla L\left({}_{S_b}\mathbf{p}\right)\right\|}. \end{aligned} \tag{6}$$

In this expression, $\mathbf{p}$ is obtained by backprojecting pixels with the network depth and $\nabla L(\cdot)$ stands for the gradient of the log-odd L, defined in (4). Depending on which frame is used for the *point*, we have either frame-to-map $\mathbf{e}_\mathrm{o2p}^{k,o(k)}$ or map-to-map $\mathbf{e}_\mathrm{o2p}^{m,o(m)}$ factors [26]. Specifically, $\mathbf{e}_\mathrm{o2p}^{k,o(k)}$ is formulated by the kth frame with its most overlapping submap $o(k)$, while $\mathbf{e}_\mathrm{o2p}^{m,o(m)}$ is defined by the mth submap with its most overlapping submap $o(m)$. In this paper, the frame-to-map factor constrains the current frame from the stereo network so that the factor can be built before the optimization, while the map-to-map factor aligns submaps using the fused depth that was integrated in the mth submap.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.

The occupancy-to-point weighting matrix with the assumption of linear occupancy near the surface is defined as

$$\mathbf{W}_{\mathrm{o2p}}^{a,b} = \mathrm{diag}\left(\begin{bmatrix} w_{\mathrm{o2p}_1} & \cdots & w_{\mathrm{o2p}_N} \end{bmatrix}\right),$$
$$w_{\mathrm{o2p}_i} = \frac{1}{\sigma_{\mathrm{map}_i}^2 + \sigma_{d_i}^2}, \quad \sigma_{\mathrm{map}_i} = \frac{l_{\mathrm{min}}}{\left\|3\nabla L({}_{S_b}\mathbf{p})\right\|}, \tag{7}$$

where the total uncertainty is the sum of the occupancy map uncertainty σ_{map} and the network depth uncertainty σ_d. l_{min} stands for the minimum log-odd value as defined in (3). In contrast to highly precise LiDAR measurements, the depth uncertainty, which often includes noisy observations, should be assigned pixel-wise to properly weight its contribution to the optimization loss. In the following section, we describe how the uncertainty is obtained from networks.

D. Uncertainty-aware depth fusion

It is pivotal to obtain a reliable dense depth as well as the associated uncertainty for our downstream tasks. Our key idea is to fuse static and motion stereo, which are complementary to each other — static stereo provides a reliable small baseline even when stationary, while motion stereo potentially brings a large baseline, depending on the camera motion. Our method does not depend on a specific network, but we found that Unimatch [13] and the MVS network [17] work well in real-world scenes. However, the previous networks only predict disparity or depth without any uncertainties. Therefore, we augment the base architecture with an uncertainty decoder and adopt the Laplacian loss function for (aleatoric) uncertainty learning. Specifically, our stereo network loss function is

$$\mathcal{L}_{\mathrm{st}}(\boldsymbol{\theta}) = \mathcal{L}_u(\boldsymbol{\theta}) + \mathcal{L}_{\nabla u_x}(\boldsymbol{\theta}) + \mathcal{L}_{\nabla u_y}(\boldsymbol{\theta}), \tag{8}$$

where $\boldsymbol{\theta}$ is the network weights, and u stands for the disparity. The disparity loss $\mathcal{L}_u$, modeled in the Laplacian distribution as in [8], [9], is defined as

$$\mathcal{L}_u(\boldsymbol{\theta}) = \sum_{i \in \mathcal{T}} \frac{\left|u_i - u_{\mathrm{gt}_i}\right|}{\sigma_{u_i}} + \log \sigma_{u_i}, \tag{9}$$

where $\mathcal{T}$ is a training set including pairs of stereo images and the ground-truth disparity. We additionally add the gradient loss $\mathcal{L}_{\nabla u}$ for sharper uncertainty output, which is analogously defined as $\mathcal{L}_u$. The gradient uncertainty along the horizontal and vertical directions is derived from the disparity uncertainty as

$$\sigma_{\nabla u_x} = \sqrt{\sigma_u^2(x+1,y) + \sigma_u^2(x-1,y)},$$
$$\sigma_{\nabla u_y} = \sqrt{\sigma_u^2(x,y+1) + \sigma_u^2(x,y-1)}. \tag{10}$$

We propagate the uncertainty from the disparity to the depth with linearization,

$$\hat{d}_{\mathrm{st}} = \frac{f_c b}{u}, \quad \sigma_{\mathrm{st}} = \frac{f_c b}{u^2}\sigma_u, \tag{11}$$

where f_c is the rectified focal length, b is the stereo baseline.

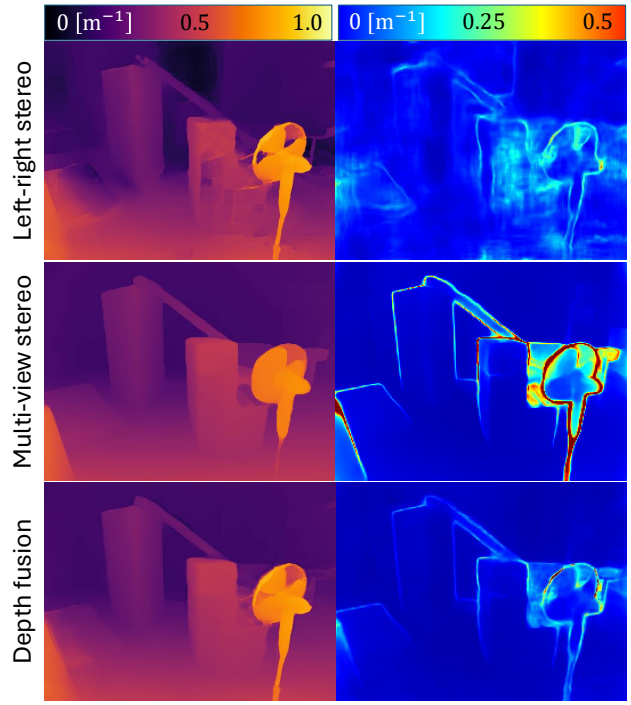

Fig. 3. Predicted inverse (only for visualization purpose) depth and its corresponding standard deviation of (top) stereo network with the 11 cm baseline, (middle) MVS network with the 50 cm maximum baseline among 8 views, and (bottom) depth fusion in the EuRoC dataset.

Likewise, we modify the loss function of the MVS network [17]

$$\mathcal{L}_{\mathrm{mvs}}(\boldsymbol{\phi}) = \sum_{i \in \mathcal{T}} \frac{\left|\log d_i - \log d_{\mathrm{gt}_i}\right|}{\sigma_{l_i}} + \log \sigma_{l_i}, \tag{12}$$

where the network learns log-depth uncertainty. We transform the log-depth to a depth with a linearized model,

$$\hat{d}_{\mathrm{mvs}} = \exp\left(\log d_i\right), \quad \sigma_{\mathrm{mvs}} = \hat{d}_{\mathrm{mvs}}\sigma_l. \tag{13}$$

Given the pixel-wise estimates from the networks $(\hat{d}_{\mathrm{st}}, \sigma_{\mathrm{st}})$, $(\hat{d}_{\mathrm{mvs}}, \sigma_{\mathrm{mvs}})$ and with the assumption that two estimates are independent, we can optimally fuse two depth estimates [28],

$$\hat{d}_{\mathrm{fuse}} = \sigma_{\mathrm{fuse}}^2\left(\sigma_{\mathrm{st}}^{-2}\hat{d}_{\mathrm{st}} + \sigma_{\mathrm{mvs}}^{-2}\hat{d}_{\mathrm{mvs}}\right),$$
$$\sigma_{\mathrm{fuse}}^2 = \left(\sigma_{\mathrm{st}}^{-2} + \sigma_{\mathrm{mvs}}^{-2}\right)^{-1}. \tag{14}$$

Fig. 3 shows an example where the fan stand has more detailed depth in the stereo network, while farther objects are sharper in the MVS network. Fused depth naturally maintains the optimal depth based on uncertainty-aware fusion.

E. Submap management

It is crucial to maintain uniformly distributed submaps with sufficient, yet not redundant overlaps between consecutive submaps for the occupancy-to-point factor. To achieve this, we keep tracking the overlap volume between the current depth frustum and the latest submap. If this falls below a certain threshold ϵ_{vol}, a new submap is generated. We

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.

TABLE I

ABSOLUTE TRAJECTORY ERROR [M] IN THE EUROC DATASET

	VINS-Fusion[1][4]	OKVIS2[2][1]	DVI-SLAM[3][31]	Ours	ORB-SLAM3[3][2]	OKVIS2[2][1]	Ours
	Causal				*Non-causal*		
MH01	0.166	0.034	0.042	0.034	0.036	0.021	**0.019**
MH02	0.152	0.029	0.046	0.031	0.033	0.020	**0.019**
MH03	0.125	0.039	0.081	0.037	0.035	0.032	**0.029**
MH04	0.280	0.065	0.072	0.072	**0.051**	0.059	0.071
MH05	0.284	0.087	0.069	0.081	0.082	**0.061**	0.066
V101	0.076	0.037	0.059	**0.035**	0.038	**0.035**	**0.035**
V102	0.069	0.031	0.034	0.028	**0.014**	**0.014**	**0.014**
V103	0.114	0.035	0.028	0.029	0.024	**0.020**	0.021
V201	0.066	0.039	0.040	0.033	0.032	**0.023**	0.024
V202	0.091	0.030	0.039	0.028	**0.014**	0.019	0.016
V203	0.096	0.041	0.055	0.041	0.024	0.027	**0.022**
Avg	0.138	0.043	0.051	0.041	0.035	**0.030**	**0.030**

[1] Results taken from [2].
[2] Results obtained ourselves.
[3] Results taken from the paper.

TABLE II

MESH ACCURACY [M] WITHOUT THRESHOLD AND COMPLETENESS [%] WITH 0.2M THRESHOLD IN THE EUROC DATASET

	Simplemapping[1][17]	Ours-Mono	Ours	Simplemapping[1][17]	Ours-Mono	Ours
	Accuracy			*Completeness*		
V101	0.176	0.076	**0.042**	40.58	41.74	**44.28**
V102	0.123	0.073	**0.041**	55.07	54.96	**58.38**
V103	0.161	0.093	**0.048**	47.74	55.92	**68.48**
V201	0.155	0.090	**0.072**	41.93	39.62	**43.64**
V202	0.156	0.073	**0.069**	56.37	54.79	**57.66**
V203	0.090	0.089	**0.072**	**62.58**	58.42	59.96
Avg	0.144	0.082	**0.057**	50.71	50.91	**55.40**

[1] Results obtained ourselves with the same metric.

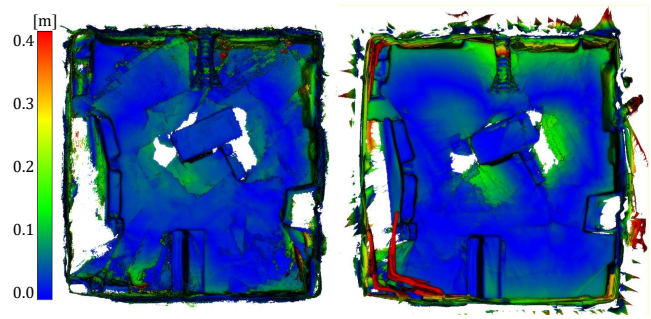

Fig. 4. Reconstructed mesh with mesh-to-point error encoding in V102 sequence. (Left) Ours and (Right) Simplemapping [17].

approximate this by sampling points in the frustum and test whether points have been observed in the previous submap.

We build the occupancy-to-point factors based on the the oldest covisible submap for geometric loop closure. We record covisible submaps to the latest submap in terms of co-observed sparse landmarks. Then, the most overlapping submap is declared as the oldest one among covisible submaps with at least M_{min} points. The submap poses are updated accordingly after optimizing the factor graph (5), in order to maintain a globally consistent map.

V. EXPERIMENTAL RESULTS

We evaluate our method on two public datasets, EuRoC [29] and Hilti-Oxford [10] in terms of localization and mapping accuracy. The estimated trajectory is aligned in $\mathrm{SE}(3)$ to the ground-truth before evaluation. We report localization accuracy as absolute trajectory error (ATE) RMSE in EuRoC or the predefined score where the position error $[1, 10]$cm is scaled to $[0, 100]$ points in the Hilti-Oxford dataset. We report causal and non-causal errors where the former only takes measurements up to the current time, while the latter is the final loop-closed trajectory [1]. To evaluate mapping accuracy, submap meshes are reconstructed using marching cubes [30] and combined using submap poses. Estimated and ground-truth vertices are downsampled with a voxel size of 1cm^3, then point-to-plane ICP aligns both sets of vertices. We report accuracy as an average distance from all estimated vertices to the ground-truth without threshold in meter and completeness as a fraction of ground-truth vertices that are within 0.2m to the estimated vertices.

A. Implementation Details

We fine-tune the stereo network [13] starting from pre-trained weights on the TartanAir dataset [32] without the post-refinement for fast inference. Then, we add a decoder adopted from MonoDepth2 [33] after the correlation volume for uncertainty learning while freezing the other weights in the loss (8). Likewise, we add an uncertainty decoder to the MVS network [17], which has the same architecture as the depth decoder of our base network. Then, we train uncertainty starting from the pre-trained weights using the loss (13) while freezing the rest of weights in the ScanNet dataset [34]. After training, we calibrate the uncertainty prediction by multiplying by a constant gain to ensure that the averaged Mahalanobis distance equals the unit distance in the validation set.

Incoming images are downsampled to a 512×384 resolution in $10\,$Hz for the networks. We input 8 images with up-to-date poses after optimizing the factor graph with a varying baseline of $3\,$cm or $10\,$cm and a sparse landmark depth map to our MVS network. The voxel resolution is set to $2.5\,$cm at each submap. We set $l_{\text{min}} = -5.015$, $\epsilon_{\text{vol}} = 0.2$, $M_{\text{min}} = 100$ and take maximum of 200 and 1000 points for the frame-to-map and map-to-map factors, respectively. We downsample the points with the lowest uncertainty if the number of points is above the threshold. We use Ceres solver [35] for the optimization and use the same parameters per dataset. All the experiments were run on a PC with an Intel i7-13700 CPU and an Nvidia RTX3080 GPU.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.

TABLE III

LOCALIZATION SCORES IN THE HILTI-OXFORD DATASET

	OKVIS2[1]	Ours	BAMF-SLAM[2][36]	XR Penguin[2][37]	OKVIS2[1]	Ours
	Causal		*Non-causal*			
exp01	8.46	4.62	15.38	40.00	45.38	**55.38**
exp02	4.09	0.45	4.55	16.82	15.91	**24.09**
exp03	0.00	0.00	8.24	**16.47**	0.00	0.00
exp07	6.67	15.00	1.67	11.67	15.00	**21.67**
exp09	5.00	2.50	1.25	15.00	10.62	**16.88**
exp11	4.00	2.00	2.00	20.00	36.00	**38.00**
exp15	14.44	20.00	7.78	24.67	**47.78**	43.33
exp21	0.00	0.00	0.00	**4.00**	0.00	0.00
Total	42.66	44.57	40.86	150.62	170.70	**199.35**

[1] Results obtained ourselves with online extrinsics calibration.
[2] Results taken from the leaderboard [38].

B. The EuRoC Dataset

This dataset provides stereo images and IMU measurements recorded by a drone, the ground-truth trajectory and point clouds of the Vicon room in mm-level accuracy [29]. We use the stereo pair for our stereo network and left images for our MVS network. Table I summarizes ATE where all competitors also use a stereo and inertial configuration. Our method improves the baseline, OKVIS2 [1] by adding occupancy-to-point factors in causal evaluation. We obtained the same accuracy on average when compared to the baseline in non-causal evaluation. This indicates that the depth fusion could not achieve 3 cm-level accuracy, which is already quite accurate. On the other hand, our approach outperforms state-of-the-art methods with nontrivial margins. Since our method tightly couples localization and volumetric mapping, we also evaluate the mapping accuracy in Table II. For fair comparison to Simplemapping [17] in a monocular-inertial set up, we implement *Ours-Mono* where only left images are used without the stereo network. The proposed method achieves the highest accuracy and completeness, while *Ours-Mono* also outperforms the competitor. Fig. 4 compares the mapping accuracy where *Ours* improves mesh accuracy especially along the edge of the room.

C. The Hilti-Oxford dataset

This dataset provides images from 5 cameras and IMU measurements captured by a handheld device, sparse ground-truth positions, and mm-accurate dense point clouds [10]. We use all 5 cameras for the visual-inertial estimator, 2 front cameras for the stereo network, and the front left camera for the MVS network. We calibrate camera-IMU extrinsic parameters online. Table III reports the localization score in the Challenge sequences where our non-causal implementation also includes the final bundle adjustment. At time of writing, our method is ranked first place on the leaderboard among the published methods. In terms of mean position error, our causal score corresponds to 13.3 cm and the non-causal score to 6.1 cm in sequences from which the score was obtained. Fig. 1 visualizes the estimated trajectory in colored submap meshes in exp04 where the ground-truth point clouds are available — our method gives 9.4cm accuracy and 62.2% completeness.

TABLE IV

ABLATION STUDY: ATE IN CAUSAL (C), NON-CAUSAL (NC), ACCURACY (ACC.), AND COMPLETENESS (COMP.) IN EUROC DATASET

	S	U	F	ATE-C [m]	ATE-NC [m]	Acc. [m]	Comp. [%]
Baseline [1]				0.043	**0.030**	-	-
+ Stereo network	✓			0.045	0.032	0.080	50.38
+ Uncertainty	✓	✓		0.043	0.031	0.072	50.31
+ Depth fusion	✓	✓	✓	**0.041**	**0.030**	**0.057**	**55.40**

TABLE V

TIMING IN MEAN$\pm\sigma$ [MS] MEASURED IN THE EUROC DATASET

	VI optimization	Occupancy update	Stereo network	MVS network
OKVIS2 [1]	24.2 ± 5.3	-	-	-
Ours	34.3 ± 6.6	11.6 ± 3.6	56.3 ± 24.1	75.9 ± 9.7

D. Ablation Study

This study shows the effectiveness of our contributions. Table IV summarizes averages over all sequences in the EuRoC dataset. The localization accuracy has been improved with uncertainty-awareness. Furthermore, we improve mapping accuracy 10% by having network-predicted uncertainty over the heuristic quadratic uncertainty [5] and 29% by the depth fusion. Fig. 1 qualitatively shows less noisy and more complete meshes of our depth fusion when compared to the stereo network with a heuristic uncertainty model.

E. Timings and memory

We measure computation time in MH01 of the EuRoC dataset as shown in Table V. The real-time estimator of our method with loop-closure optimizations running in the background can run in 29 Hz, while OKVIS2 as our baseline can run in 41 Hz. Our mapping can run at 13 Hz given the bottleneck of the MVS network. Also, MVS and stereo networks only take 3.51GB of GPU memory during inference.

VI. CONCLUSION

We have proposed visual-inertial SLAM that tightly couples localization and volumetric occupancy mapping through uncertainty-aware occupancy-to-point factors. To further improve the mapping accuracy, we fine-tune stereo and MVS networks with uncertainty learning. Then, we optimally fuse depths from these networks based on their predicted uncertainty, which is employed in the optimization and depth integration in submapping. Through evaluation in benchmark datasets, our method outperforms the state-of-the-art regarding accuracy both in localization and mapping, while providing dense occupancy in real-time (13Hz). For future work, we aim to consider dynamic objects, incorporate epistemic depth uncertainty, and close the loop with robot navigation and control accounting for the state uncertainty.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] S. Leutenegger, "OKVIS2: Realtime Scalable Visual-Inertial SLAM with Loop Closure," *arXiv preprint arXiv:2202.09199*, 2022.

[2] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[3] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4666–4672.

[4] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," *arXiv preprint arXiv:1901.03642*, 2019.

[5] N. Funk, J. Tarrio, S. Papatheodorou, M. Popović, P. F. Alcantarilla, and S. Leutenegger, "Multi-resolution 3D mapping with explicit free space representation for fast and accurate mobile robot motion planning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3553–3560, 2021.

[6] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *IEEE International Symposium on Mixed and Augmented Reality*, 2011, pp. 127–136.

[7] F. Furrer, T. Novkovic, M. Fehr, A. Gawel, M. Grinvald, T. Sattler, R. Siegwart, and J. Nieto, "Incremental Object Database: Building 3D Models from Multiple Partial Observations," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 6835–6842.

[8] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3227–3237.

[10] L. Zhang, M. Helmberger, L. F. T. Fu, D. Wisth, M. Camurri, D. Scaramuzza, and M. Fallon, "Hilti-oxford dataset: A millimeter-accurate benchmark for simultaneous localization and mapping," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 408–415, 2023.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, pp. 7–42, 2002.

[12] G. Xu, Y. Wang, J. Cheng, J. Tang, and X. Yang, "Accurate and efficient stereo matching via attention concatenation volume," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2461–2474, 2023.

[13] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 941–13 958, 2023.

[14] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[16] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman, and C. Godard, "SimpleRecon: 3D reconstruction without 3D convolutions," in *the European Conference on Computer Vision (ECCV)*, 2022, pp. 1–19.

[17] Y. Xin, X. Zuo, D. Lu, and S. Leutenegger, "Simplemapping: Real-time visual-inertial dense mapping with deep multi-view stereo," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2023, pp. 273–282.

[18] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1697–1716, 2016.

[19] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM: Real-time dense monocular slam with learned depth prediction," in *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6243–6252.

[20] T. Laidlow, J. Czarnowski, and S. Leutenegger, "DeepFusion: Real-time dense 3D reconstruction for monocular SLAM using single-view depth and gradient predictions," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4068–4074.

[21] V. Reijgwart, A. Millane, H. Oleynikova, R. Siegwart, C. Cadena, and J. Nieto, "Voxgraph: Globally consistent, volumetric mapping using signed distance function submaps," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 227–234, 2019.

[22] L. Koestler, N. Yang, N. Zeller, and D. Cremers, "TANDEM: Tracking and dense mapping in real-time using deep multi-view stereo," in *Conference on Robot Learning*, 2022, pp. 34–45.

[23] A. Rosinol, J. J. Leonard, and L. Carlone, "Probabilistic volumetric fusion for dense monocular SLAM," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3097–3105.

[24] Z. Teed and J. Deng, "Droid-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[25] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1281–1292.

[26] S. Boche, S. B. Laina, and S. Leutenegger, "Tightly-Coupled LiDAR-Visual-Inertial SLAM and Large-Scale Volumetric Occupancy Mapping," *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[27] S. B. Laina, S. Boche, S. Papatheodorou, D. Tzoumanikas, S. Schaefer, H. Chen, and S. Leutenegger, "Scalable Autonomous Drone Flight in the Forest with Visual-Inertial SLAM and Dense Submaps Built without LiDAR," *arXiv preprint arXiv:2403.09596*, 2024.

[28] N. A. Carlson, "Federated square root filter for decentralized parallel processors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 26, no. 3, pp. 517–525, 1990.

[29] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[30] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.

[31] X. Peng, Z. Liu, W. Li, P. Tan, S. Cho, and Q. Wang, "DVI-SLAM: A dual visual inertial SLAM network," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12 020–12 026, 2024.

[32] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4909–4916.

[33] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 3828–3838.

[34] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5828–5839.

[35] S. Agarwal, K. Mierle, and Others, "Ceres Solver," 8 2024. [Online]. Available: https://github.com/ceres-solver/ceres-solver

[36] W. Zhang, S. Wang, X. Dong, R. Guo, and N. Haala, "BAMF-SLAM: Bundle adjusted multi-fisheye visual-inertial slam using recurrent field transforms," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6232–6238.

[37] Y. Wang, Y. Ng, I. Sa, A. Parra, C. Rodriguez-Opazo, T. Lin, and H. Li, "MAVIS: Multi-Camera Augmented Visual-Inertial SLAM using SE 2 (3) Based Exact IMU Pre-integration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 1694–1700.

[38] "HILTI SLAM Challenge 2022 Leaderboard," Feb. 2025. [Online]. Available: https://hilti-challenge.com/leader-board-2022.html

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 02:54:18 UTC from IEEE Xplore. Restrictions apply.