TES-CVIDS: A Transmission Efficient Sub-Map Based Collaborative Dense VI-SLAM Framework

Tianjun Zhang, Lin Zhang, *Senior Member, IEEE*, Fengyi Zhang, Shengjie Zhao, *Senior Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—In recent years, how to achieve stable localization and construct high-quality dense maps in large-scale scenes has become a research highlight. In large-scale scenes, for the consideration of the mapping accuracy and efficiency, multi-agent systems rather than single-agent ones are usually employed. Currently, as far as we know, collaborative VI-SLAM (Visual Inertial Simultaneous Localization And Mapping) systems applicable to multi-agent systems are still sporadic, and systems those can achieve a good balance among the localization accuracy, the mapping density, and the transmission efficiency are temporarily lacking. In this paper, we propose a novel centralized collaborative VI-SLAM framework, namely TES-CVIDS (Transmission Efficient Sub-map based Collaborative Visual-Inertial Dense SLAM). In TES-CVIDS, instead of the original RGBD images, the compact sub-maps are transmitted, effectively reducing the transmission data redundancy. After that, the server completes key-frame processing, hierarchical pose-graph optimization, and global dense map construction in three separate threads. Besides, thanks to our depth search mechanism, the geometry information of all key-frames can be recovered on the server-end. Thus, sub-maps can be regenerated after the global pose-graph optimization to maintain the consistency between the localization and the mapping. Both the qualitative and the quantitative experimental results corroborate the superior performance of our TES-CVIDS.**

***Index Terms*—Multi-agent, transmission efficient, dense mapping, visual-inertial odometry.**

I. Introduction

IN NUMEROUS automation fields, ranging from augmented reality [1], [2] to autonomous driving [3], [4], an accurate and profound understanding of the surrounding environment is typically essential. As a reliable solution to achieve such an understanding, SLAM (Simultaneous Localization And Mapping) has made significant progress over the past decade or so [5], [6], [7]. Among all of the research sub-branches in this field, visual-inertial SLAM, VI-SLAM in short, has drawn increasing interests of researchers in recent years [8], [9], [10] due to its stable tracking performance and cost-effective sensor configurations. Currently, VI-SLAM systems have been widely integrated to UAVs (Unmanned Aerial Vehicle), wheeled robots, smartphones and many other types of devices. Unfortunately, since the transformations among the reference coordinate systems of different agents cannot be obtained, most of these systems are inadequate for the multi-agent systems without complex extensions. In reality, the robots or devices are sometimes put into use in the form of clusters or formations, such as drone formations and wearable motion capture systems. In such cases, these single-agent-oriented systems no longer work, and collaborative SLAM frameworks designed for the multi-agent systems must be employed instead.

Different from the standard single-agent-oriented SLAM systems, to share and exchange the information necessary for localization and mapping, the communication among multiple agents are essential for collaborative SLAM systems. Actually, most existing collaborative SLAM systems are sparse, meaning that only the sparse feature points and corresponding descriptors need to be transmitted. Thus, the bandwidth pressures in such systems can typically be ignored. In recent years, so as to support decision-making tasks like navigation and obstacle avoidance, SLAM systems are often required to be able to yield dense maps rather than just sparse feature maps in both academia and industry. As a sub-branch, collaborative SLAM systems are naturally pinned on similar expectations.

In dense collaborative SLAM systems, the volume of data that needs to be shared among different agents far exceeds that in sparse systems, resulting in a significant increase in bandwidth pressure. Currently, existing dense collaborative SLAM systems mostly select to transmit RGBD images of key-frames directly or after a straightforward image compression. However, such transmission modes create great bandwidth pressure. Even if images are compressed, the improvement is usually a drop in the bucket. Aside from these frameworks that transmit images directly, there is also some work that transmits sub-maps instead of utilizing the image based transmission. Such a mode can effectively eliminate the inter-frame data redundancy and reduce

Manuscript received 21 January 2024; revised 29 March 2024; accepted 2 April 2024. Date of publication 8 April 2024; date of current version 27 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62272343 and Grant 61936014, in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Lin Zhang.)*

Tianjun Zhang, Lin Zhang, Fengyi Zhang, and Shengjie Zhao are with the School of Software Engineering, Tongji University, Shanghai 201804, China, and also with the Engineering Research Center of Key Software Technologies for Smart City Perception and Planning, Ministry of Education, Shanghai 201804, China (e-mail: 1911036@tongji.edu.cn; cslinzhang@tongji.edu.cn; 2131507@tongji.edu.cn; shengjiezhao@tongji.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@um.edu.mo).

To make our results reproducible, the source code has been released at https://cslinzhang.github.io/TES-CVIDS-MainPage/.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3385452.

Digital Object Identifier 10.1109/TIV.2024.3385452

the bandwidth pressure. Nevertheless, since original RGBD images are not shared among different agents, updating the shared sub-maps without additional communication costs when the poses of key-frames change after global optimization becomes quite challenging. Therefore, these schemes usually assume that the relative poses among key-frames in the same sub-map are absolutely accurate, and only optimize the relative poses among different sub-maps during the global pose optimization, which undoubtedly influences the localization accuracy of the system.

Thus far, as far as we know, most of the current existing collaborative VI-SLAM systems cannot fully satisfy the requirements in localization accuracy, mapping density and transmission efficiency simultaneously. As an attempt to fill in the aforementioned research gaps, we propose a novel collaborative dense VI-SLAM system, namely TES-CVIDS. Our contributions can be mainly summarized as follows.

1) A novel collaborative dense VI-SLAM system, TES-CVIDS, is proposed, which follows a centralized client-server architecture. To improve the communication efficiency between all clients and the central server in TES-CVIDS, the sub-map based transmission instead of the image based one is used, which significantly relieves the pressure on the transmission bandwidth of the system. The server disperses the tasks of key-frame processing, pose-graph optimization and global dense map construction into three different threads and the latter two tasks are both running in backend threads, guaranteeing the real-time processing of TES-CVIDS.
2) A novel space-efficient and outlier-aware probabilistic TSDF (Truncated Signed Distance Function) sub-map representation is designed. In our proposed representation, the 3D space is firstly divided into thousands of chunks, and occupied chunks are further divided into quantities of voxels, which greatly saves the storage space compared to the fully voxel-based representation. Besides, we model the distributions of the TSDF values both stored in voxels and corresponding to depth observations, enabling the probabilistic incremental updates of our sub-map representation and ensuring its robustness to outlier observations.
3) An efficient two-stage segmented pose-graph optimization pipeline is proposed and integrated to TES-CVIDS. In our pipeline, a skeleton sub-graph of the global pose graph to be optimized is extracted, and the remaining frames are naturally divided into multiple sequences. In the first stage of the pipeline, only the skeleton sub-graph is optimized by non-linear optimization. Afterward, the poses of key-frames in remaining sequences are updated by an efficient EM-based (Expectation-Maximum) smooth algorithm in the second stage. By integrating our novel pose-graph optimization pipeline, TES-CVIDS achieves a superior localization performance in both the speed and the accuracy.
4) A key-frame depth search strategy with adaptive step sizes is proposed. The strategy simulates the rendering process and can recover both the depth structure and the color information of key-frames from sub-maps efficiently. The search processes of all pixels in the recovered RGBD images are executed in parallel, and the step lengths are determined adaptively to ensure efficient convergence. When the relative poses among key-frames in a sub-map change significantly after the global pose-graph optimization, the depth search processes for these key-frames are activated. Then the sub-map is updated according to the new poses of key-frames to ensure the consistency between the localization results and the global dense maps.

To make the reported results in this paper fully reproducible, we have publicly released our source code and data at https://cslinzhang.github.io/TES-CVIDS-MainPage/.

II. Related Work

A. Dense V-SLAM Systems

In 2011, Newcombe et al. proposed a RGBD dense SLAM framework, namely KinectFusion [11]. In KinectFusion, ICP (Iterative Closest Point) [12] is employed to align the global map and the point cloud of the current frame scanned by the RGBD sensor for tracking. In [13], Endres et al. proposed a feature point based RGBD SLAM system, RGBDSLAMv2. Instead of matching the point cloud of different frames, Endres et al. selects to extract sparse features from RGB images and then matches them, which is a commonly utilized idea in monocular SLAM systems. Both of these two frameworks are able to achieve real-time dense reconstruction, but as pioneering work, their accuracy and stability in localization and mapping are not satisfactory. To improve localization accuracy, Concha and Civera proposed RGBDTAM [14], which introduces the visual direct method [15] and the photometric error to RGBD SLAM. However, such a design also brings some negative effects to the system, such as the sensitivity to changes of light conditions. In 2017, a milestone work, namely BundleFusion [16], was presented. In BundleFusion, tracking is conducted in a coarse-to-fine way. Initially, the poses of key-frames are obtained via the sparse feature matching, and then the coarse estimation is further refined by jointly minimizing both the photometric error and the point-to-plane geometry error. In [17], Sun et al. proposed Plane-Edge-SLAM, in which both surfaces and edges are extracted from the RGB images with the assistance of the depth information. After the extraction, the surfaces and the edges are utilized to generate the corresponding geometry constraints in the pose estimation process so as to compensate for the lack of visual features. Different from the aforementioned schemes, BAD-SLAM [18] presented by Schops et al. selects to use the surfel map rather than the TSDF or point cloud one. In BAD-SLAM, the 3D points captured by the RGBD camera are used to minimize geometry distances between them and their corresponding nearest surfels in the map. This approach enables the simultaneous adjustment of the key-frames' poses and the positions of surfels. In 2019, Shan et al. [19] designed a dense VI-SLAM system based on VINS-Mono [9]. By using the RGBD camera, in addition to the dense mapping capabilities, both initialization and tracking in their system become much more stable compared to VINS-Mono [9].

Compared with the dense mapping based on RGBD cameras, owing to the affordable manufacturing cost and the lightweight structure of the sensor, the problem of monocular dense mapping in an online manner has also garnered significant research interest in the past decade. Many remarkable work, such as [20], [21], [22], [23], have been presented. However, compared with RGBD-based or stereo-based systems, the mapping stability and accuracy of monocular systems are usually obviously inferior. Fortunately, with the rapid development in deep learning, significant progress has been made in the field of monocular depth estimation in recent years [24], [25], [26], [27]. By integrating these approaches, RGBD SLAM systems can usually run successfully even if only a monocular camera is used. Thus, currently, how to obtain depth maps, whether through hardware or algorithms, is relatively unimportant for RGBD SLAM systems.

B. Collaborative V-SLAM Systems

The earliest collaborative V-SLAM system can be traced back to the work of Forster et al. in [28], which followed a traditional structure-from-motion pipeline. As a seminal work, its implementations are relatively straightforward and have obvious limitations in both accuracy and robustness. In the same year, another milestone collaborative SLAM system namely CoSLAM [29] was proposed. In CoSLAM, the influence on the localization accuracy brought by dynamic objects is considered, and the map points are classified into static points and dynamic ones. CoSLAM demonstrates superior robustness to the dynamic environment. However, in CoSLAM, all agents are required to be strictly time-synchronized, which undoubtedly increases the hardware cost and limits its application scopes. In 2016, Deutsch et al. creatively abstracted the client-end of the collaborative SLAM system and proposed a loosely coupled framework [30]. In such a system, the odometry running on each agent is regarded as a black box, ignoring detailed inner implementations. Such a mechanism brings an outstanding hardware adaptation ability to the system. Nevertheless, since it is a 2D SLAM system, the current advanced 3D visual odometries can't be integrated with it. CCM-SLAM [31] presented by Schmuck et al. is a tightly-coupled centralized monocular collaborative SLAM system with outstanding localization accuracy. It possesses a relatively modern architecture and shows superior localization performance. In [32], a fully distributed SLAM system, namely DOOR-SLAM, was proposed, which is based on peer-to-peer communication and does not require full connectivity among the robots. In recent years, many researchers have focused on upgrading sensors from monocular cameras to visual-inertial camera suites to improve localization stability, and some remarkable work has been published [8], [9], [10]. However, it's a pity that most of these systems only support the single-agent mode. To address this issue, CVI-SLAM [33] was developed as a collaborative SLAM system specifically designed for monocular visual-inertial suites, following a similar design as CCM-SLAM [31], but with the equipped sensor on each agent replaced by a visual-inertial suite. The upgraded sensors significantly improve the final localization accuracy of CVI-SLAM [33].

Apart from the aforementioned sparse systems, there are also some dense ones. C ^{2} TAM [34] proposed by Riazuelo et al. is a cloud framework for collaborative SLAM based on RGBD cameras. Golodetz et al. also proposed a collaborative RGBD SLAM system in [35]. It supports multiple users to interactively reconstruct dense voxel-based models of the large-scale environment. In [36], the first fully distributed multi-robot system for dense metric-semantic SLAM, namely Kimera-Multi, was presented. Kimera-Multi is capable of building accurate 3D metric-semantic meshes while being robust to incorrect loop closures and requiring less computation than other distributed SLAM backends. CVIDS [37] proposed by Zhang et al. is the first collaborative visual-inertial SLAM framework that supports dense mapping without the depth sensor. It exhibits superior accuracy performance in both the localization and the mapping. The aforementioned dense schemes can usually achieve accurate localization and high-quality mapping. However, all of them require real-time transmission of RGBD images from the clients to the server, which puts significant pressure on network bandwidth.

III. System Overview

The overall framework of TES-CVIDS is illustrated in Fig. 1. The client-end of TES-CVIDS runs on each agent and can theoretically be integrated with any existing VIO (Visual-Inertial Odometry). On each agent, 2D features and corresponding descriptors, sparse 3D map points and poses in the local reference coordinate system of key-frames are packed and then sent to the central server for collaborative localization. Besides, the RGB images captured by the equipped camera, the poses yielded by the local VIO and the depth maps (either captured directly or estimated by algorithms) are fed to the local mapping module to construct a local sub-map in our proposed space-efficient and outlier-aware probabilistic TSDF form. More details about our sub-map representation can be found in Section IV. Once the volume of the current sub-map achieves the preset threshold, it is converted to a compact data package and then transmitted to the central server. After that, a new sub-map is constructed.

The communication module on the server-end of TES-CVIDS is mainly responsible for unpacking the data received from the client-end of all agents, including both key-frame messages and sub-map messages. Once unpacked, the data is then fed to the co-localization module and the global mapping module. For these two modules, we will introduce them in detail in Sections V and VI, respectively.

IV. Compact Sub-Map Representations

In the client-end of TES-CVIDS, both RGB images and depth maps collected by each agent over a period of time are encoded into the sub-map using our proposed compact representations before the transmission, so as to eliminate the inter-frame information redundancy and relieve the bandwidth pressure. To achieve this goal, we model both the distribution of the TSDF value in each voxel and the observation model of the depth values with respect to the corresponding voxels' TSDF values in a probabilistic way. Besides, we also offer the recursive state

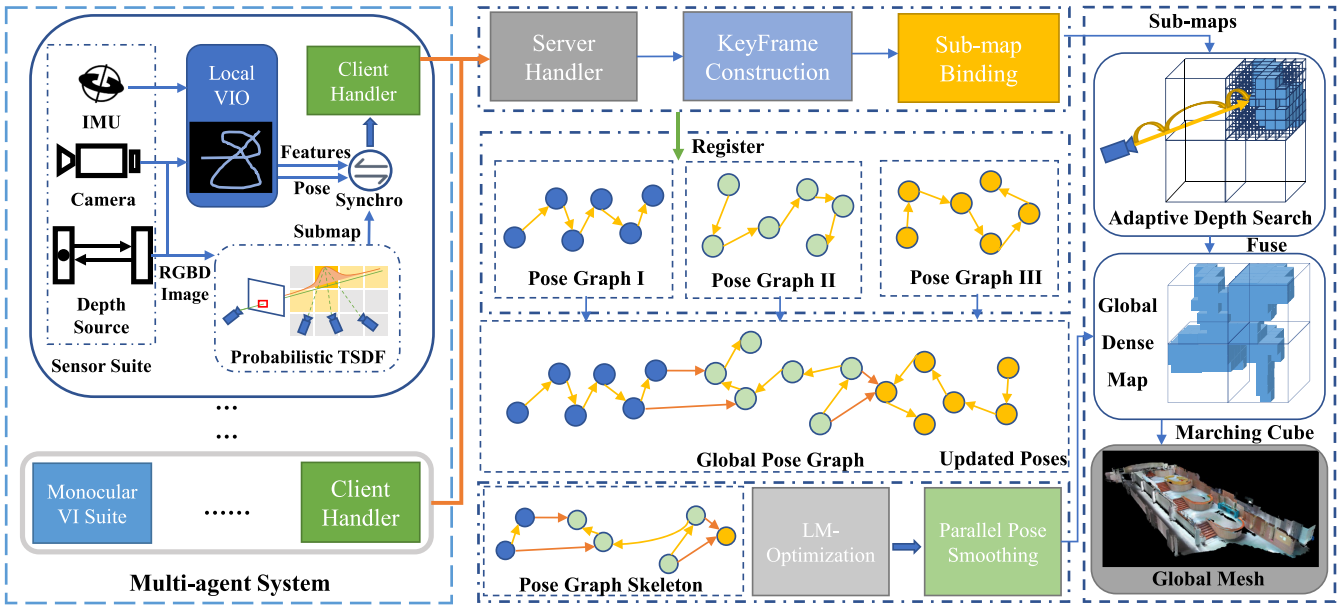


Fig. 1. Overall architecture of TES-CVIDS. Features, poses of key-frames and sub-maps are packed and then sent to the server from the client by the client handler. Then the co-localization module aligns the local pose graph of each agent to the global one and conduct global pose-graph optimization using our two stage pipeline. Besides, the dense mapping module runs concurrently in a separate thread to fuse the received sub-maps to the global one. Finally, via the marching cube process, the dense structure of the scene can be represented in the mesh form.

update equation, which is used when new RGBD frames are obtained. Therefore, our proposed sub-map representation can be updated incrementally just like the standard TSDF one.

A. Outlier-Aware Probabilistic TSDF Sub-Map

Given a sequence of frames $\{\mathcal{F}^1, \ldots, \mathcal{F}^r\}$, the TSDF value $\tilde{\tau}_{\boldsymbol{v}}$ of a voxel $\boldsymbol{v}$ observable in these frames can be determined by corresponding TSDF observations $\tau_{\boldsymbol{v}}^1, \ldots, \tau_{\boldsymbol{v}}^r$ of these frames. Furthermore, regardless of how the depth map and corresponding TSDF values are obtained (i.e., through sensors or algorithms), outliers are bound to occur. For instance, stereo-based depth estimations usually perform inadequately in weak-texture regions, while ToF depth cameras often fail to yield usable depths for the edges and corners of observed objects. In view of this, we use $\rho_{\boldsymbol{v}}$ to represent the probability to get inlier TSDF observations at $\boldsymbol{v}$. Assuming independent observations, the posteriori of $\tilde{\tau}_{\boldsymbol{v}}$ and $\rho_{\boldsymbol{v}}$ is given as,

$$p(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}} | \tau_{\boldsymbol{v}}^1, \ldots, \tau_{\boldsymbol{v}}^r) \propto p(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}) \prod_{i=1}^{r} p(\tau_{\boldsymbol{v}}^i | \tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}), \quad (1)$$

where $p(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}})$ is the prior distribution. For ease of representation, the posteriori $p(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}} | \tau_{\boldsymbol{v}}^1, \ldots, \tau_{\boldsymbol{v}}^i)$ is represented as $p_i(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}})$. Then the recurrence relationship of the posteriori can be expressed as,

$$p_i(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}) \propto p_{i-1}(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}) p(\tau_{\boldsymbol{v}}^i | \tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}). \quad (2)$$

Motivated by the compact representation proposed in [38], the likelihood probability distribution $p(\tau_{\boldsymbol{v}}^i | \tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}})$ can be modeled as a combination of a normal distribution (representing inlier measurements) and a uniform one (representing outlier measurements), which can be expressed as,

$$p(\tau_{\boldsymbol{v}}^i | \tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}) = \rho_{\boldsymbol{v}} \mathcal{N}(\tau_{\boldsymbol{v}}^i | \tilde{\tau}_{\boldsymbol{v}}, ({\wr_{\boldsymbol{v}}}^i)^2) + (1 - \rho_{\boldsymbol{v}}) \mathcal{U}(\tau_{\boldsymbol{v}}^i | \tau_{\min}, \tau_{\max}), \quad (3)$$

where $\tilde{\tau}_{\boldsymbol{v}}$ and ${\wr_{\boldsymbol{v}}}^i$ are the expectation and the standard deviation of the observation model, respectively, and $\tau_{\min}$ ($\tau_{\max}$) stands for the minimum (maximum) possible TSDF value. In our implementations, $\tau_{\min}$ and $\tau_{\max}$ are set to $-\tau_{truc}$ and τ_{truc}, respectively, where τ_{truc} is the truncation distance of the map. How to determine $\tilde{\tau}_{\boldsymbol{v}}$ and ${\wr_{\boldsymbol{v}}}^i$ will be discussed in Section IV-B.

Equation (2) depicts the recurrence relationship of the posteriori distribution. Unfortunately, multiplying by a Gaussian distribution, the Gaussian-uniform mixture distribution does not keep its original form. Thus, some approximations are necessary to achieve an incremental update. Specifically, the posteriori in (2) is approximated by the product of a Beta distribution and a normal one, which is formulated as,

$$p_i(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}}) \approx q(\tilde{\tau}_{\boldsymbol{v}}, \rho_{\boldsymbol{v}} | a_{\boldsymbol{v}}^i, b_{\boldsymbol{v}}^i, \mu_{\boldsymbol{v}}^i, \sigma_{\boldsymbol{v}}^i) = Beta_{\boldsymbol{v}}^i \mathcal{N}_{\boldsymbol{v}}^i, \quad (4)$$

where $a_{\boldsymbol{v}}^i$ and $b_{\boldsymbol{v}}^i$ controls the Beta distribution $Beta_{\boldsymbol{v}}^i$, and $\mu_{\boldsymbol{v}}^i$ and $\sigma_{\boldsymbol{v}}^i$ are the expectation and the stadard deviation of the Gaussian distribution $\mathcal{N}_{\boldsymbol{v}}^i$, respectively. For each time a new observation is received, the parameters in the posteriori alter, including $a_{\boldsymbol{v}}^i$, $b_{\boldsymbol{v}}^i$, $\mu_{\boldsymbol{v}}^i$ and $\sigma_{\boldsymbol{v}}^i$, but the form of the distribution remains, which allows to update the posteriori incrementally. It's worth mentioning that, the inlier measurement ratio $\rho_{\boldsymbol{v}}$ can be determined by $a_{\boldsymbol{v}}^i$ and $b_{\boldsymbol{v}}^i$ since they control the Beta distribution $Beta_{\boldsymbol{v}}^i = Beta(\rho_{\boldsymbol{v}} | a_{\boldsymbol{v}}^i, b_{\boldsymbol{v}}^i)$.

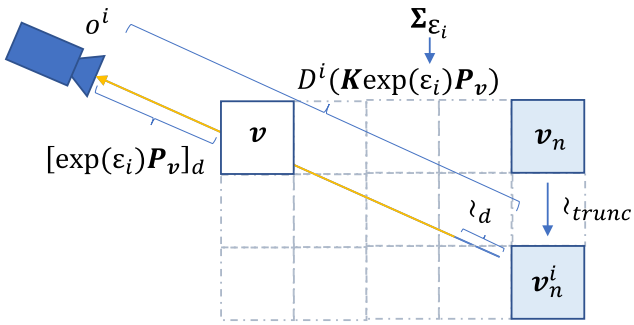


Fig. 2. Illustration of the TSDF observation model adopted in TES-CVIDS. There are mainly three parts of the "inaccuracy" in our observation model, including the inaccuracy of the camera pose, the inaccuracy of the depth sensor and the inaccuracy brought by approximating $\boldsymbol{v}_n$ to $\boldsymbol{v}_n^i$, corresponding to $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^i}$, $\wr_d$ and $\wr_{trunc}$ in the figure, respectively.

B. Observation Model of TSDF Values

In this subsection, we will introduce how to determine $\tilde{\tau}_{\boldsymbol{v}}$ and $\wr_{\boldsymbol{v}}^i$ utilized in (3), which are the expectation and the standard deviation of TSDF values in our observation model, respectively. The illustration of our observation model is given in Fig. 2. For a voxel $\boldsymbol{v}$, its nearest occupied voxel on the ray from the sensor origin $\boldsymbol{o}^i$ to $\boldsymbol{v}$ is defined as $\boldsymbol{v}_n^i$, and $\boldsymbol{v}$'s nearest occupied voxel is defined as $\boldsymbol{v}_n$. In our observation model, a valid and necessary assumption is that, the expectation of $\boldsymbol{v}_n^i$ is just $\boldsymbol{v}_n$. Thus, we have,

$$\tau_{\boldsymbol{v}} = f(\boldsymbol{D}, \boldsymbol{T}) = \boldsymbol{D}(\boldsymbol{K}\boldsymbol{T}\boldsymbol{P}_{\boldsymbol{v}}) - [\boldsymbol{T}\boldsymbol{P}_{\boldsymbol{v}}]_d, \tag{5}$$

where $\boldsymbol{K}$ and $\boldsymbol{T}$ are the intrinsics matrix and the pose of the camera, respectively, $\boldsymbol{D}$ is the depth map, and $\boldsymbol{P}_{\boldsymbol{v}}$ is the 3D coordinate of $\boldsymbol{v}$ and symbol $[*]_d$ stands for the depth value (coordinate in z axis) of the inner 3D point. Then $\tilde{\tau}_{\boldsymbol{v}}$ can be considered as the signed distance $\tilde{\tau}_{\boldsymbol{v}}^i$ between $\boldsymbol{v}$ and $\boldsymbol{v}_n^i$, and thus we have,

$$\tilde{\tau}_{\boldsymbol{v}} = \tilde{\tau}_{\boldsymbol{v}}^i = f(\boldsymbol{D}^i, \boldsymbol{T}^i), \tag{6}$$

where $\boldsymbol{D}^i$ and $\boldsymbol{T}^i$ are the depth map and the pose of the camera in $\mathcal{F}^i$, respectively.

Next, we will define the standard deviation $\wr_{\boldsymbol{v}}^i$ of our observation model. We consider that the inaccuracy of the measurement $\tau_{\boldsymbol{v}}^i$ mainly consists of three parts, the inaccuracy of pose $\boldsymbol{T}^i$, the inaccuracy of the depth sensor and the inaccuracy brought by approximating $\boldsymbol{v}_n$ to $\boldsymbol{v}_n^i$. Then, $\wr_{\boldsymbol{v}}^i$ can be modeled as,

$$(\wr_{\boldsymbol{v}}^i)^2 = (\boldsymbol{J}_{\boldsymbol{\varepsilon}^i}^{\tilde{\tau}_{\boldsymbol{v}}})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^i}(\boldsymbol{J}_{\boldsymbol{\varepsilon}^i}^{\tilde{\tau}_{\boldsymbol{v}}})^T + (\wr_d)^2 + (\wr_{trunc})^2, \tag{7}$$

where $\boldsymbol{J}_{\boldsymbol{\varepsilon}^i}^{\tilde{\tau}_{\boldsymbol{v}}}$ is the jacobian of $\tilde{\tau}_{\boldsymbol{v}}$ to the pose vector $\boldsymbol{\varepsilon}^i$ corresponding to $\boldsymbol{T}^i$ in Lie algebra form, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^i}$ is the convariance matrix of $\boldsymbol{\varepsilon}^i$, $\wr_d$ and $\wr_{trunc}$ are the standard deviation of the Gaussian white noise brought by the defect of the "depth sensor" and by approximating $\tilde{\tau}_{\boldsymbol{v}}$ to $\tilde{\tau}_{\boldsymbol{v}}^i$, respectively. It's worth mentioning that, the "depth sensor" here is a broad concept and it can also represent any depth estimation algorithm. In our implementations, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}^i}$ is initialized to a constant matrix according to the accuracy of the local odometry, and $\wr_d$ and $\wr_{trunc}$ are proportional to the depth measurement and the absolute value of the signed distance, respectively. As for the jacobian $\boldsymbol{J}_{\boldsymbol{\varepsilon}^i}^{\tilde{\tau}_{\boldsymbol{v}}}$, it can be obtained using the chain rule as,

$$\boldsymbol{J}_{\boldsymbol{\varepsilon}^i}^{\tilde{\tau}_{\boldsymbol{v}}} = \frac{\partial \tilde{\tau}_{\boldsymbol{v}}}{\partial \boldsymbol{D}^i} \cdot \frac{\partial \boldsymbol{D}^i}{\partial \boldsymbol{p}_{\boldsymbol{v}}^{iT}} \cdot \frac{\partial \boldsymbol{p}_{\boldsymbol{v}}^i}{\partial (\boldsymbol{\varepsilon}^i)^T}, \tag{8}$$

where $\boldsymbol{p}_{\boldsymbol{v}}^i$ is the projection of voxel $\boldsymbol{v}$ on the depth map $\boldsymbol{D}^i$. By approximating the second term $\partial \boldsymbol{D}^i / \partial \boldsymbol{p}_{\boldsymbol{v}}^{iT}$ with the intensity gradient of $\boldsymbol{D}^i$ at $\boldsymbol{p}_{\boldsymbol{v}}^i$, which can be computed by the Sobel operator, the jacobian $\boldsymbol{J}_{\boldsymbol{\varepsilon}^i}^{\tilde{\tau}_{\boldsymbol{v}}}$ can be easily obtained. Till now, we have deduced the final form of the standard deviation $\wr_{\boldsymbol{v}}{}^i$ of the observation model.

C. Chunk-Wise Voxel Management

To support scalable reconstruction, voxels are not managed directly in TES-CVIDS. Instead, the reconstructed scene is evenly divided into small chunks, with each chunk consisting of a cluster of adjacent voxels. Mathematically, a chunk can be considered as a set of voxels, that is, chunk $\boldsymbol{C}_{\boldsymbol{P}}$ located at $\boldsymbol{P} = [X_{\boldsymbol{P}}, Y_{\boldsymbol{P}}, Z_{\boldsymbol{P}}]^T$ with an edge length of k can be represented as,

$$\boldsymbol{C}_{\boldsymbol{P}} = \{\boldsymbol{v}_{\boldsymbol{P}'} | \boldsymbol{P} \preceq \boldsymbol{P}' \prec \boldsymbol{P} + \boldsymbol{S}\}, \tag{9}$$

where $\boldsymbol{P}'$ is the position of voxel $\boldsymbol{v}_{\boldsymbol{P}'}$, and $\boldsymbol{S}$ is the size of chunk $\boldsymbol{C}_{\boldsymbol{P}}$ which can be represented as,

$$\boldsymbol{S} = [ks_{\boldsymbol{v}}, ks_{\boldsymbol{v}}, ks_{\boldsymbol{v}}]^T, \tag{10}$$

where $s_{\boldsymbol{v}}$ is the resolution of the sub-map (the edge length of a voxel). Thus, it's easy to know that a chunk consists of k^3 voxels in sum. In TES-CVIDS, the operations of create, modify and search are all initially conducted at the chunk level and then at the voxel level, which effectively improves the computational efficiency and reduces the usage of the storage.

D. Submap Encoding

Before being sent to the server, the sub-map is encoded into a compact representation so as to improve the transmission efficiency. First, the voxels with a low inlier ratio $\rho_{\boldsymbol{v}}$ are eliminated. In our implementations, the ratio threshold is set to 0.5. After the screening, the sub-map is converted from our proposed probabilistic form to the standard TSDF one. Specifically, by defining the voxel before and after the conversion as $\boldsymbol{v}$ and $\boldsymbol{v}_t$, the TSDF value $\tau_{\boldsymbol{v}_t}$ of $\boldsymbol{v}_t$ is just the discrete sampling of $\boldsymbol{v}$'s TSDF expectation $\mu_{\boldsymbol{v}}$. After the discrete sampling, $\tau_{\boldsymbol{v}_t}$ can be stored in an integer form rather than the floating point one. In our implementations, the sampling step is set to $1e-4$, which equals to 0.1 mm. Since only reliable voxels remain after eliminating high-outlier-ratio ones, $\boldsymbol{v}_t$'s weight is set to be constant and won't be transmitted.

Except for the aforementioned basic encoding pipeline, TES-CVIDS also supports the compression of the sub-map based on tensor-train [39], which can further effectively compress the encoded sub-map. As a mature tensor compression scheme, it will not be discussed in detail in this paper. It's worth mentioning that tensor-train can achieve a high compression ratio theoretically, but it is a form of lossy compression. Therefore, we recommend

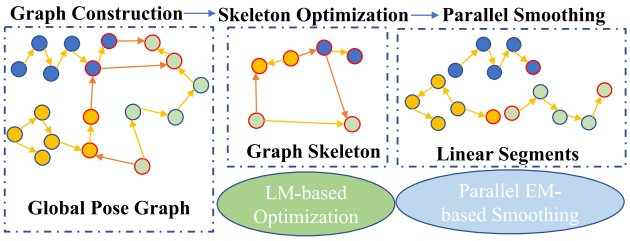


Fig. 3. Illustration of our proposed two-stage segmented pose-graph optimization pipeline utilized in TES-CVIDS. In the first stage, the LM-based optimization is used to optimize the skeleton of the pose-graph, and in the second stage, our proposed EM-based smoothing is adopted to optimize the remaining frames in parallel.

utilizing such compression only if the available bandwidth is insufficient.

V. Co-Localization Module

The co-localization module of TES-CVIDS incorporates several mechanisms from CVIDS [37], including the construction and registration of key-frames, the BoW-based loop closure detection (LCD), the alignment of local reference coordinate systems (CSs), the determination of data associations in the global pose graph, and the four-DoF pairwise consistency evaluation of all loop closure measurements. Necessary engineering modifications and extensions have been made to adapt these mechanisms to TES-CVIDS. Besides, to further enhance both the accuracy and the efficiency of the localization, we further propose a novel two-stage segmented pose-graph optimization pipeline and integrate it to TES-CVIDS. In the first stage of our pipeline, a skeleton sub-graph of the global graph is extracted and then optimized. After that, in the second stage the remaining key-frames are assigned to quantities of segments and their poses are updated via our EM-based pose smoothing in parallel. The illustration of the our proposed two-stage pose-graph optimization pipeline is given as Fig. 3.

A. Skeleton Extraction of the Pose Graph

In TES-CVIDS, it is somewhat challenging to efficiently achieve the convergence by optimizing the large-scale global pose-graph directly using the non-linear optimization. Fortunately, the local VIO of each agent can yield relatively accurate short-term pose estimations in most cases, errors in the global pose graph mainly accumulate at nodes associated with loop closures or where the local VIO performs unsatisfactorily. As an attempt to improve the speed and the accuracy simultaneously, in the first stage of our pose-graph optimization pipeline, we extract a skeleton sub-graph of the global pose graph and just optimize the poses in this sub-graph rather the entire global graph. Before extracting the skeleton, some special key-frames which compose the basic structure of the skeleton sub-graph should be selected firstly, which are,

1) The key-frames that are connected to others in loop closures.
2) The key-frames that are connected to others with inaccurate associations.

Since each data association connects two key-frames, the selected key-frames can be considered as a quantity of key-frame pairs. As for how to distinguish inaccurate data associations, it will be discussed in Section V-B. After selecting the basic structure, each pair of the selected key-frames, their adjacent key-frames and adjacent key-frames of adjacent key-frames form a segment of the skeleton. If two segments have common key-frames, they are further merged.

B. Discrimination of Inaccurate Data Associations

Existing VIOs can achieve stable short-term tracking in most cases. However, at some special moments, there may be obvious localization errors due to light changes, motion blur, IMU jamming, etc., which ultimately introduce inaccurate data associations. If the global pose graph is directly optimized, these inaccurate associations will obviously affect the overall localization accuracy. Therefore, we propose a discriminant index specifically designed for VIOs to effectively identify these inaccurate associations for targeted processing later.

Currently, mainstream VIOs usually solve the poses by utilizing both visual and inertial information in a joint framework. However, in reality, one source of the information sometimes may be noisy, resulting in an inaccurate pose estimation. Therefore, the pose estimation of a key-frame should be confident enough and consistent with both the visual information and the inertial one. Otherwise, the data association between this key-frame and its former one can be considered inaccurate. In TES-CVIDS, two adjacent key-frames are considered to be connected inaccurately if the latter key-frame $\mathcal{F}^i$ satisfies any of the following criteria,

1) No enough visual data are used in its pose estimation.
2) Pose estimation is inconsistent with visual observations.
3) The inertial data used in pose estimation is insufficient.

Next, we will explain these three criteria in detail one by one.

Since it's both complex and time-consuming to compute the specific position distribution of each map point, the distributions of all map points are assumed to be the same fixed one. Hence, for criterion 1), it is sufficient to check if the number of map points that are visible for $\mathcal{F}^i$ is less than the threshold t_{mp}. Similarly, criterion 2) equals to check if the average reprojection error of each point is less than the threshold t_{rep}. For the last criterion which pertains to the inertial data, we build the indicator based on the preintegration volume $\mathcal{I}^{i-1,i}$ between $\mathcal{F}^i$ and its previous key-frame $\mathcal{F}^{i-1}$, and the mathematical form of criterion 3) can be given as,

$$Log(\|\mathbf{\Sigma}_{ine}^{i-1,i}\|_F) < t_{ine}, \tag{11}$$

where $\mathbf{\Sigma}_{ine}^{i-1,i}$ is the covariance matrix corresponding to $\mathcal{I}^{i-1,i}$, t_{ine} is the threshold and $\|*\|_F$ stands for the Frobenius norm. The Frobenius norm of the covariance matrix usually varies exponentially. That is why we convert it to the log domain.

Through quantitative evaluations, we find among all criteria, criterion 3) is the most important. Thus, in TES-CVIDS, we use fixed and relatively loose thresholds for the first two criteria

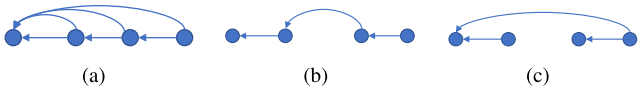


Fig. 4. Sketch maps of three different connection modes in the pose-graph optimization pipeline of TES-CVIDS. The sketches of the dense connection, the sparse connection and the skip connection are displayed from (a) to (c).

and an adaptive one for criterion 3). Specifically, t_{mp} and t_{rep} are set to 10 and 4.0, respectively, and t_{ine} can be adaptively determined as,

$$t_{ine} = Avg(Log_{ine}) + \sigma_{ine}, \tag{12}$$

where $Avg(Log_{ine})$ is the average value of the inertial indicator of all key-frames in the global pose graph and σ_{ine} is the corresponding standard deviation.

C. Pose-Graph Optimization of the Skeleton Sub-Graph

Before the optimization of the skeleton sub-graph, the connection relationships among the nodes in the sub-graph need to be established. In this regard, three types of connection modes are defined, namely dense connection, sparse connection and skip connection, which are illustrated in Fig. 4. For key-frames in the same segment without inaccurate data associations, they are connected using dense connections, which means that each key-frame is associated with sequential edges to its five previous key-frames, so as to maintain the accurate local structure of the global pose graph. Each segment of the skeleton is just connected to its adjacent two segments in the sparse connection mode. Besides, in the segment with inaccurate associations, the skip connections are used. In the skip connection mode, inaccurate data associations are ignored and sequential edges do not "stride across" these associations. Instead, considering two key-frames connected by the inaccurate association as a pair, a sequential edge between the previous frame to the former one of the pair and the subsequent frame to the latter one is added. Once the nodes and the connections of the skeleton sub-graph are determined, the poses of corresponding key-frames can be optimized using any non-linear optimization scheme. In TES-CVIDS, the LM (Levenberg-Marquadt) scheme [40] is adopted.

D. Parallel Pose Smoothing

As aforementioned, once the skeleton sub-graph is extracted and optimized, the remaining key-frames can be naturally divided to multiple segments, and each segment can be efficiently updated in parallel. Since the graph in each segment follows a simple linear structure, it is somewhat time-consuming to use the non-linear optimization. Instead, we resort to an EM-based pose smoothing scheme, which is also utilized in [37]. Taking a segment $\boldsymbol{\mathcal{S}}^j$ as an example, which can be represented as,

$$\boldsymbol{\mathcal{S}}^j = \{\mathcal{F}^i | I^j < i < I^j + N^j\}, \tag{13}$$

where I^j is the index of the first key-frame in $\boldsymbol{\mathcal{S}}^j$ and N^j is the number of key-frames in the segment. In $\boldsymbol{\mathcal{S}}^j$, we select to use the dense connection mode mentioned in Section V-C. Besides, except for the key-frames in $\boldsymbol{\mathcal{S}}^j$, the poses of key-frames $\mathcal{F}^{I^j-1}$ and $\mathcal{F}^{I^j+N^j+1}$ are also involved in the pose smoothing, while their poses are set to be fixed and won't be updated. The pose smoothing problem amounts to,

$$\min_{\boldsymbol{\mathcal{T}}^j} \sum_{i=I^j}^{I^j+N^j+1} \sum_{(i,k)\in\boldsymbol{\mathcal{E}}^i} ||\boldsymbol{e}(\boldsymbol{T}_i, \boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik})||_2^2, \tag{14}$$

where $\boldsymbol{\mathcal{E}}^i$ is the set of key-frames' indices that are connected to $\mathcal{F}^i$, $\boldsymbol{T}_i$ ($\boldsymbol{T}_k$) stands for the pose of the keyframe $\mathcal{F}_i$ ($\mathcal{F}_k$), $\hat{\boldsymbol{T}}_{ik}$ is the constraint of the relative pose between $\mathcal{F}_i$ and $\mathcal{F}_k$, $\boldsymbol{\mathcal{T}}^j$ represents all poses of corresponding key-frames in $\boldsymbol{\mathcal{S}}^j$ and $\boldsymbol{e}(\boldsymbol{T}_i, \boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik})$ is the four-DoF (the yaw angle and the translation) error that can be defined as,

$$\begin{aligned} \boldsymbol{e}\left(\boldsymbol{T}_i, \boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik}\right) &= \left[e_{ik}^{yaw}, \left(\boldsymbol{e}_{ik}^{\boldsymbol{t}}\right)^T\right]^T \\ \boldsymbol{e}_{ik}^{yaw} &= \theta_k^{yaw} - \theta_i^{yaw} - \hat{\theta}_{ik}^{yaw} \\ \boldsymbol{e}_{ik}^{\boldsymbol{t}} &= \boldsymbol{R}_i\left(\boldsymbol{t}_k - \boldsymbol{t}_i\right) - \hat{\boldsymbol{t}}_{ik}, \end{aligned} \tag{15}$$

where θ_i^{yaw}, θ_k^{yaw} and $\hat{\theta}_{ik}^{yaw}$ are corresponding yaw angles of poses $\boldsymbol{T}_i$, $\boldsymbol{T}_k$ and $\hat{\boldsymbol{T}}_{ik}$, respectively, $\boldsymbol{R}_i$ is the rotation matrix of $\boldsymbol{T}_i$, and $\boldsymbol{t}_i$, $\boldsymbol{t}_k$ and $\hat{\boldsymbol{t}}_{ik}$ are translation vectors of corresponding poses. For the pose smoothing process, an important inequality is given as,

$$||\boldsymbol{e}(\boldsymbol{T}_i, \boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik})||_2^2 \leq \frac{1}{2}(||\boldsymbol{e}_i(\boldsymbol{T}_i, \hat{\boldsymbol{T}}_{ik})||_2^2 + ||\boldsymbol{e}_k(\boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik})||_2^2), \tag{16}$$

where $\boldsymbol{e}_i(\boldsymbol{T}_i, \hat{\boldsymbol{T}}_{ik})$ and $\boldsymbol{e}_k(\boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik})$ are defined as,

$$\begin{aligned} \boldsymbol{e}_i(\boldsymbol{T}_i, \hat{\boldsymbol{T}}_{ik}) &= \left[\theta_i^{yaw} + \hat{\theta}_{ik}^{yaw} - {}_E\hat{\theta}_{ik}^{yaw}, \boldsymbol{R}_i\left(\boldsymbol{t}_i - {}_E\hat{\boldsymbol{t}}_{ik}\right) - \hat{\boldsymbol{t}}_{ik}\right]^T \\ \boldsymbol{e}_k(\boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik}) &= \left[\theta_k^{yaw} - {}_E\hat{\theta}_{ik}^{yaw}, \boldsymbol{R}_i(\boldsymbol{t}_k - {}_E\hat{\boldsymbol{t}}_{ik})\right]^T, \end{aligned} \tag{17}$$

where ${}_E\hat{\theta}_{ik}^{yaw}$ and ${}_E\hat{\boldsymbol{t}}_{ik}$ can be any constant. For ease of representation, we use $\boldsymbol{e}_i$ and $\boldsymbol{e}_k$ to represent $\boldsymbol{e}_i(\boldsymbol{T}_i, \hat{\boldsymbol{T}}_{ik})$ and $\boldsymbol{e}_k(\boldsymbol{T}_k, \hat{\boldsymbol{T}}_{ik})$, respectively. According to (16), an approximated version of (14) can be obtained as,

$$\min_{\boldsymbol{\mathcal{T}}^j} \sum_{i=I^j}^{I^j+N^j+1} \sum_{(i,k)\in\boldsymbol{\mathcal{E}}^i} \left(||\boldsymbol{e}_i||_2^2 + ||\boldsymbol{e}_k||_2^2\right). \tag{18}$$

It can be easily proved that, the optimal solutions of (14) and (18) will be the same when,

$$\begin{aligned} {}_E\hat{\theta}_{ik}^{yaw} &= \left(\tilde{\theta}_i^{yaw} + \tilde{\theta}_k^{yaw} - \hat{\theta}_{ik}^{yaw}\right)/2 \\ {}_E\hat{\boldsymbol{t}}_{ik} &= \left(\tilde{\boldsymbol{t}}_i + \tilde{\boldsymbol{t}}_k - \boldsymbol{R}_i^T\hat{\boldsymbol{t}}_{ik}\right)/2, \end{aligned} \tag{19}$$

where $\tilde{\theta}_i^{yaw}$, $\tilde{\theta}_k^{yaw}$, $\tilde{\boldsymbol{t}}_i$ and $\tilde{\boldsymbol{t}}_k$ are all optimal solutions of (14). Since these optimal solutions are unavailable, the EM framework [41] is adopted to smooth all poses iteratively. In the E-step, we utilize the current values of all frame poses to compute ${}_E\hat{\theta}_{ik}^{yaw}$ and ${}_E\hat{\boldsymbol{t}}_{ik}$. Then in the M-step, since each error term in (18) is only related to the pose of one frame, we can obtain the analytical optimal solution and then update the corresponding pose. By updating the poses of all keyframes and the approximated

optimal solutions alternately, the smoothed poses will converge efficiently.

VI. Global Dense Mapping Module

Once the communication module on the server-end receives and unpacks the sub-map sent by a client, the sub-map is added to the task list of the global dense mapping module. In such a module, a separate back-end global mapping thread is running to perform the global dense mapping task. During each loop of the thread, all sub-maps in the task list are traversed, and sub-maps who have already been aligned to the global coordinate system are linked to corresponding key-frames. Then these sub-maps can be fused to the global map. Furthermore, the global poses of key-frames may change after the global pose-graph optimization. In such a case, sub-maps whose inner structure differ significantly from the initial state are regenerated to ensure the consistency between mapping and localization. Since the RGBD images are not transmitted to the central server, they are recovered based on the sub-map and the poses of key-frames via our proposed adaptive-step depth search process.

A. Sub-map Fusion

The sub-map of a client can be fused to the global TSDF map as long as the local reference coordinate system of the client has been aligned with the global one. The sub-map is transformed into the global coordinate system, and then the corresponding voxels in the global map are updated. Specifically, for a voxel $\boldsymbol{v}_{\boldsymbol{P}_s}$ located at $\boldsymbol{P}_s$ in the local coordinate system of the sub-map, its 3D position $\boldsymbol{P}_g$ in the global coordinate system can be obtained according to the transformation matriz from the local coordinate system to the global one. Then, in each of the neighbouring eight voxels of $\boldsymbol{P}_g$ in the global map, the TSDF value of $\boldsymbol{v}_{\boldsymbol{P}_s}$ and the distance between $\boldsymbol{P}_g$ and the center of the voxel are stored. Finally, the TSDF values of the corresponding voxels in the global map are updated using trilinear interpolation.

B. Global Map Regeneration

Once the global pose-graph optimization is activated, the global map should be adjusted accordingly. If no more than 25% of the sub-maps need to be updated, these sub-maps will be eliminated from the global map via inverse fusion and re-fused to the global map. Otherwise, the global map will be regenerated by all sub-maps. Regarding how to update a sub-map, there are three cases. If both the pose of the sub-map and its inner pose structure change little, the sub-map will not be updated. If the pose of the sub-map changes significantly while its inner structure remains unchanged, the content of the sub-map will not be updated, but the sub-map will need to be re-fused into the global map according to the new pose. If the inner structure of the sub-map undergoes significant changes, the sub-map will be regenerated according to the updated poses and the RGBD maps recovered by our depth search scheme, which will be discussed in Section VI-C. In TES-CVIDS, we use the pose of the first key-frame in the sub-map to represent the pose of the sub-map, and use the relative poses between the first key-frame and multiple uniformly sampled latter key-frames to model its inner structure.

Algorithm 1: Determination of DNCs.

Input: A set $\boldsymbol{\mathcal{C}}$ comprising all chunks stored in the sub-map, a dictionary $\boldsymbol{\mathcal{O}}$ comprising the occupancy states of these chunks.
Output: A map $\boldsymbol{\mathcal{D}}$ representing the DNCs of all chunks, where $\boldsymbol{\mathcal{D}}(\boldsymbol{C})$ is the DNC of chunk $\boldsymbol{C}$.

for chunk $\boldsymbol{C}$ in $\boldsymbol{\mathcal{C}}$ **do**
  $\boldsymbol{\mathcal{D}}(\boldsymbol{C}) \leftarrow DNC_{MAX}$
end for
Construct Queue $\boldsymbol{\mathcal{Q}}$
for chunk $\boldsymbol{C}$ in $\boldsymbol{\mathcal{C}}$ **do**
  Add ($\boldsymbol{C}$, 0) to $\boldsymbol{\mathcal{Q}}$ if $\boldsymbol{\mathcal{O}}(\boldsymbol{C})$ is true
end for
while $\boldsymbol{\mathcal{Q}}$ is not empty **do**
  Pop $\boldsymbol{\mathcal{Q}}$ to get the pair ($\boldsymbol{C}$, d)
  if $\boldsymbol{\mathcal{D}}(\boldsymbol{C}) > d$ **then**
    $\boldsymbol{\mathcal{D}}(\boldsymbol{C}) \leftarrow d$
    for chunk $\boldsymbol{C}_n$ in all neighbour chunks of $\boldsymbol{C}$ **do**
      Add ($\boldsymbol{C}_n, d+1$) to $\boldsymbol{\mathcal{Q}}$
    end for
  end if
end while
return $\boldsymbol{\mathcal{D}}$

C. Depth Search With Adaptive Step Length

With the sub-map represented in the TSDF form and the pose of a key-frame in this sub-map, the corresponding RGBD maps of the key-frame can be efficiently recovered via a depth search process. Specifically, for each pixel $\boldsymbol{p}$ on the map, an epipolar ray can be cast from the sensor origin $\boldsymbol{o}$ to the normalized 3D point $\boldsymbol{P}_n$ corresponding to $\boldsymbol{p}$. An ideal solution to recover RGBD maps is searching along the ray until the first occupied voxel $\boldsymbol{v}_o$ is encountered. Then, the depth of the key-frame at $\boldsymbol{p}$ can be set to the Euclidean distance between $\boldsymbol{o}$ and the position of $\boldsymbol{v}_o$, and the corresponding color can also be found in $\boldsymbol{v}_o$. Unfortunately, although the searching process of each pixel can be run in parallel, it is still time-consuming. Thus, instead of utilizing a voxel-by-voxel searching scheme, an adaptive-step searching strategy is used in our TES-CVIDS. This strategy involves conducting the search first at the chunk level and then at the voxel level, using an adaptive step length. Before the depth search, some preparations are necessary. As discussed in Section IV-C, TES-CVIDS manages voxels in a chunk-wise manner. If a chunk is not empty and contains some voxels, it can be considered as an "occupied chunk". Under such a definition, for each chunk an integer d is assigned to store the DNC (Distance to Nearest occupied Chunk) of the chunk, which indicates the number of chunks between its nearest occupied chunk and itself. For example, the DNC of an occupied chunk is assigned 0, while the DNCs of its twenty-six neighboring chunks which are not occupied are assigned 1. The pseudo-code

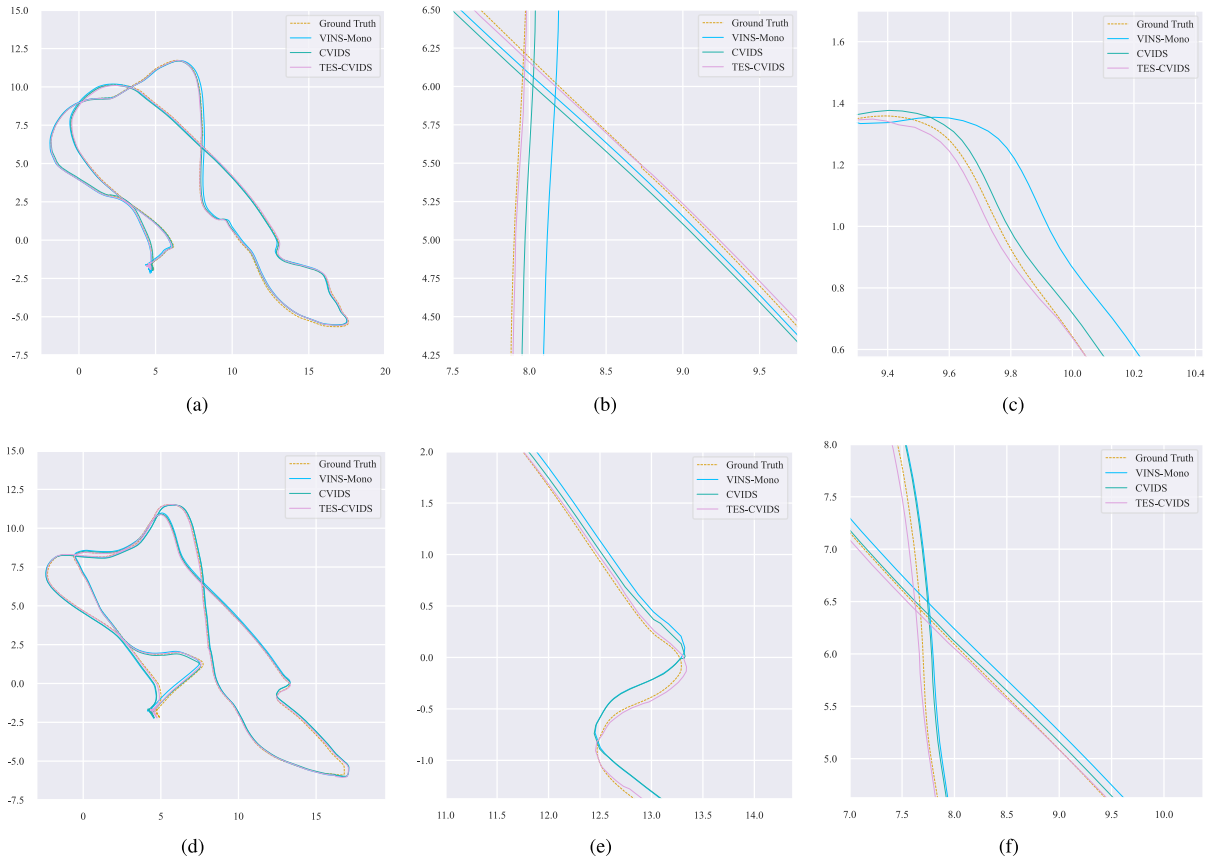


Fig. 5. Localization results of TES-CVIDS and other two counterparts (CVIDS [37] and VINS-Mono [9]) on Sequence 4 and Sequence 5 of the Euroc MH dataset. (a) shows the results on Sequence (4) and (d) corresponds to Sequence 5. To show the differences of trajectories more clearly, two enlarged local regions of (a) are offered in (b) and (c), while (e) and (f) are the enlarged regions of (d).

for determining the DNCs of all chunks is provided in Algorithm 1.

After obtaining the DNCs of all chunks, the chunk-level search is activated to find the next occupied chunk through which the search line passes. A search point $\boldsymbol{P}_s$ is initialized at the origin $\boldsymbol{O}$ of the camera and moves along the search line as the search depth increases. Assuming that the search point currently lies in chunk $\boldsymbol{C}$, the search step is set to the value for the search point to "enter" the next $d_{\boldsymbol{C}}$-th chunk, where $d_{\boldsymbol{C}}$ is the DNC of $\boldsymbol{C}$. Once the search point enters an occupied chunk, the search is executed at the voxel-level. The TSDF values of the voxels are then utilized as the guidance information to determine the step length of the search. When the search point lies in voxel $\boldsymbol{v}$, the search step $s_{\boldsymbol{v}}$ is set to,

$$s_{\boldsymbol{v}} = \max\left(|d_{\boldsymbol{v}}| - \frac{1}{2}d_{\min}, d_{\min}\right), \tag{20}$$

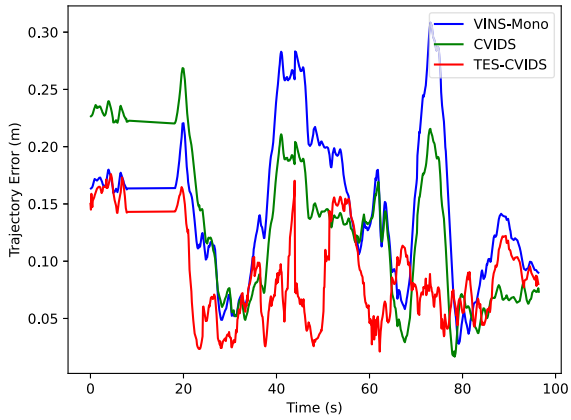
where $d_{\boldsymbol{v}}$ is the TSDF value of $\boldsymbol{v}$ and $d_{\min}$ is the edge length of the voxel. Defining voxels whose absolute TSDF values and neighbouring voxels' ones are all lower than $2d_{\min}$ as occupied voxels, the voxel-level search is conducted until the search point lies in an occupied voxel or enters the next chunk. In the former case, the depth search process of pixel $\boldsymbol{p}$ is terminated, and the corresponding RGBD values can be obtained. In the latter case, if the new chunk the search point enters is occupied, the voxel-level search continues. Otherwise, the chunk-level search is conducted instead. It is worth mentioning that if the current searched distance is larger than the maximum valid distance, the search is also terminated, and the RGBD values of $\boldsymbol{p}$ are considered unrecoverable.

VII. Experimental Results

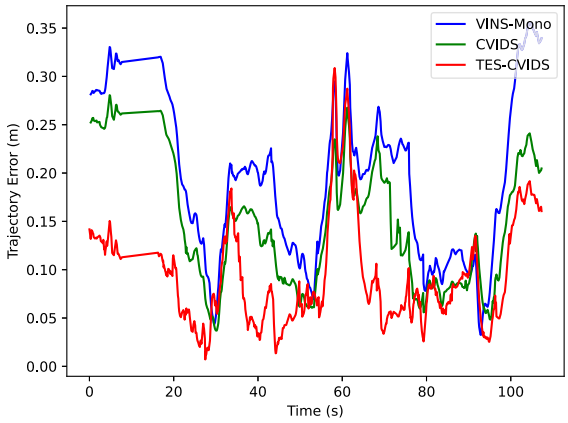
A. Experimental Setup

We evaluated the performance of our system mainly in two aspects, the localization and the mapping. On the aspect of the localization, Euroc MH (Machine Hall) dataset [42] was used for the evaluations, which includes accurate trajectory ground-truth obtained by motion capture systems. In this dataset, there are five sequences collected by a micro air vehicle, which is equipped with a global shutter camera of Aptina MT9V034, and a six-axis IMU of MEMS ADIS16448. The five sequences can be regarded as collected by five different agents. RMSE (Root Mean Squared Error) [43] was taken as the metric, which can be given as,

$$e_{RMSE} = \left(\frac{1}{M}\sum_{i=1}^{M}||trans(\mathbf{Q}_i^{-1}\mathbf{P}_i)||^2\right)^{\frac{1}{2}}, \mathbf{Q}_i, \mathbf{P}_i \in SE(3), \tag{21}$$



(a)



(b)

Fig. 6. Absolute trajectory errors along with time of TES-CVIDS and other two counterparts (CVIDS [37] and VINS-Mono [9]) on Sequence 4 and Sequence 5 of the Euroc MH dataset. (a) shows the results on Sequence (4) and (b) corresponds to Sequence 5.

where $trans(.)$ represents the translation part of the pose, M is the number of all frames, and $\{\mathbf{Q}_i\}_{i=1}^{M}$ and $\{\mathbf{P}_i\}_{i=1}^{M}$ are ground-truth and estimated poses of all frames, respectively. While on the mapping aspect, it was not proper to also use the Euroc MH dataset [42] since it only includes gray-scale images, and neither RGB images nor depth-maps are provided. Therefore, to show the mapping results and the corresponding bandwidth costs of TES-CVIDS, three groups of RGBD images and the corresponding inertial data in different scenes were collected by us using the Azure Kinect DK camera suites. In each group of data, four sequences were included, collected by four different agents. It's worth mentioning that, for the evaluation on the Euroc MH dataset in which depth maps are not offered, VINS-Mono [9] was utilized as the local odometry on the client end, while for the evaluation on our own collected dataset the VINS-RGBD [19] was chosen for better robustness and stability.

B. Qualitative Experimental Results

Collaborative Localization Effect: To qualitatively demonstrate the localization performance of TES-CVIDS, the localization results of TES-CVIDS on two "difficult" sequences of the Euroc MH dataset [42], Seuqence 4 and 5, were given in Fig. 5, and the absolute trajectory errors along with time on these two sequences were offered in Fig. 6. For comparison, the corresponding results of CVIDS [37] and VINS-Mono [9] were also provided. From Figs. 5 and 6, it can be seen that, the yielded trajectories of TES-CVIDS are much more consistent with the ground-truth ones compared with VINS-Mono [9] and CVIDS [37]. This indicates that TES-CVIDS exhibits a superior localization accuracy, which is attributed to the integration of our proposed hierarchical pose-graph optimization pipeline. For more details about the quantitative evaluation, please refer to Section VII-C.

Typical Sub-maps and Global Dense Maps: To evaluate the mapping performance of TES-CVIDS, we collected data from three different scenes using the Azure Kinect DK camera suites, which were handheld or carried by wheeled robots, and then fed the RGBD images and the inertial data to TES-CVIDS to reconstruct the scenes densely in an online manner. The final results were presented in Fig. 7, where the mapping results of the three different scenes were shown from top to bottom. Besides, for each scene, we also selected a typical sub-map to clearly display the details of the reconstruction results. From Fig. 7, it is evident that high-quality dense reconstruction results in large-scale scenes can be obtained, indicating the superior mapping accuracy performance of our TES-CVIDS.

TABLE I
RMSEs OF COMPARED VISUAL-INERTIAL LOCALIZATION SCHEMES ON EUROC MH DATASET (cm)

	VINS-Mono	VIORB	CVI-SLAM	CVIDS	**TES-C**
MH_01	12.00	7.50	8.50	3.86	**3.71**
MH_02	12.00	8.40	6.30	3.64	**3.57**
MH_03	13.00	8.70	**6.50**	7.01	7.48
MH_04	18.00	21.70	29.30	12.92	**9.14**
MH_05	21.00	8.20	**8.10**	14.53	10.37
Weighted Average	14.40	10.39	10.91	7.48	**6.69**

C. Quantitative Experimental Results

Collaborative Localization Accuracy: The quantitative evaluation of the collaborative localization accuracy of TES-CVIDS was conducted using the Euroc MH dataset [42]. In evaluation, TES-CVIDS ran under the five-agent configuration, and both RGB images and inertial data from each sequence of the dataset (from MH_01 to MH_05) were fed to a single agent. The estimated trajectories of all agents were recorded, and the corresponding RMSEs were computed and provided in Table I. Besides, in Table I, the RMSE results of other main competitors, including VINS-Mono [9], VIORB [10], CVI-SLAM [33] and CVIDS [37], were also offered. From the table, it can be observed that TES-CVIDS exhibits the lowest weighted average RMSE of all sequences among all competitors, corroborating that our TES-CVIDS possesses an outstanding localization performance.

Network Traffic: Three sequences of our collected data (Sequence 1∼3 as shown in Fig. 7(a)∼(c), respectively) were used to analyze the network traffic between the agents and the server

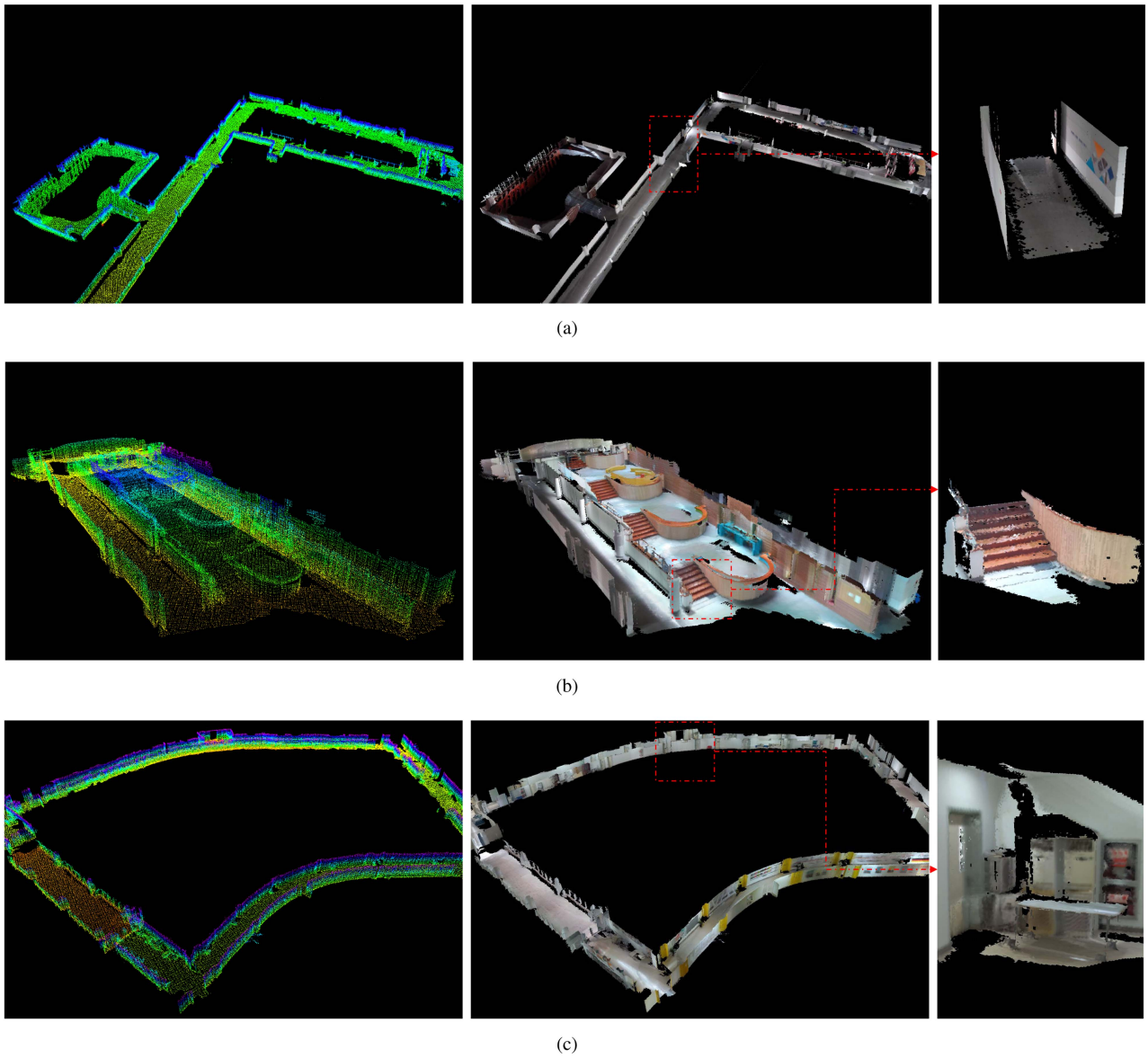


Fig. 7. Typical samples of reconstructed dense maps. From (a) to (c), the results correspond to three different scenes. And in each group, the yielded point cloud map of TES-CVIDS is shown on the left, the reconstructed mesh is given in the middle and a typical sub-map is offered on the right.

in TES-CVIDS. The collected data were fed to TES-CVIDS, which employs sub-map based data transmission, and the required average bandwidth cost to complete collaborative localization and dense reconstruction was recorded. For comparison, the image based transmission was used as a baseline. Besides, since TES-CVIDS integrates the tensor-train based sub-map compression module, the network traffic of TES-CVIDS activating such a compression module was also evaluated. Fig. 8 shows the network traffic of these three transmission modes between the server and the agents over the runtime. From Fig. 8, it can be seen that TES-CVIDS exhibits much lower bandwidth cost owing to the sub-map based transmission strategy employed compared to using the conventional image based transmission method, implying the better flexibility and robustness to the network condition. Additionally, by activating the tensor-train based compression module, the bandwidth cost of TES-CVIDS can be significantly further reduced.

Time Cost Analysis: Since the performance of the client-end of TES-CVIDS is determined by the utilized odometry, we mainly focused on analyzing the time costs of the server-end thread by thread. First, we evaluated the speed performance of the main thread and the global mapping thread. In the main thread of TES-CVIDS, the average time cost of TES-CVIDS to complete the tracking and the loop closure detection is about 29.7 ms, and fusing the sub-map to the global one in the global mapping thead takes approximately 20.2 ms per frame. Since the global mapping thread runs in parallel with the main thread, and both the depth recovery and the sub-map recomposition are not always required, TES-CVIDS can achieve a frame rate of more than 30 fps. Next, we offer the speed evaluation results

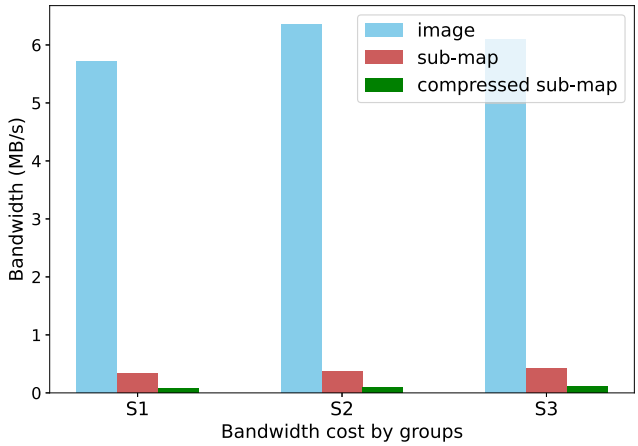


Fig. 8. Bandwidth costs of TES-CVIDS under a four-agent mode in different scenes using different transmission modes, including image based transmission, sub-map based transmission and compressed sub-map based transmission using tensor train.

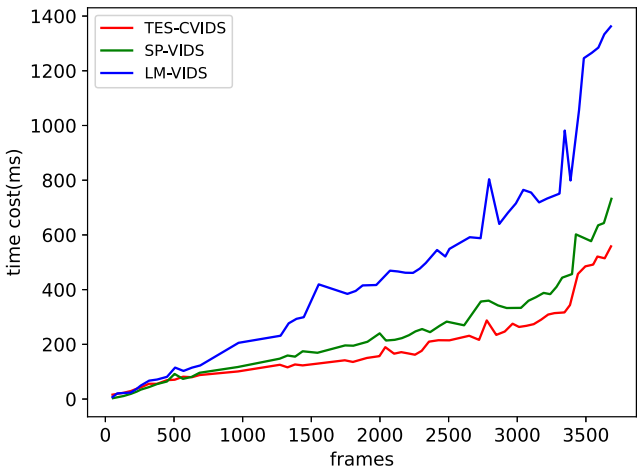
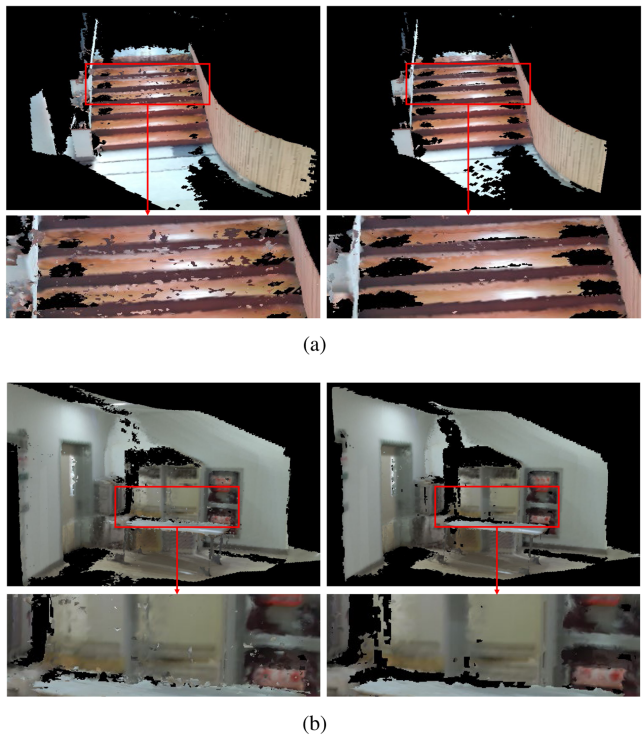


Fig. 9. Time costs of our two-stage pose-graph optimization pipeline and its other two compared variants using different numbers of involved frames.

of the localization thread, in which our two-stage pose-graph optimization pipeline is integrated. Specifically, we recorded the time consumption to complete the global optimization of the pose-graph consisting of different number of frames, since the time consumption is directly related to the number of optimized variables involved in the optimization. The results are plotted as the red curve in Fig. 9. From Fig. 9, it can be observed that TES-CVIDS can complete the global pose-graph optimization within 0.6 s even if up to 3,500 frames are involved. It is therefore evident that TES-CVIDS exhibits outstanding efficiency in both the localization and the mapping.

D. Ablation Study

Ablation Study for the Localization: The localization performance of TES-CVIDS is mainly reflected in two aspects: speed and accuracy. To verify the superior performance of our current localization configurations, we compared TES-CVIDS with other two baseline variants using both the RMSEs on each



(a)

(b)

Fig. 10. Typical samples of sub-maps represented by our proposed probabilistic TSDF form and the standard TSDF form. In each group of the results, the dense mesh extracted from the probabilistic TSDF sub-map is shown on the left, while the result corresponds to the standard TSDF map is on the right. Besides, for each group, locally enlarged regions are shown on the bottom.

TABLE II
RMSEs OF OUR SCHEME ON EUROC MH DATASET UNDER DIFFERENT LOCALIZATION CONFIGURATIONS (cm)

	SP-VIDS	LM-VIDS	**TES-CVIDS**
MH_01	4.36	5.52	**3.71**
MH_02	3.98	5.24	**3.57**
MH_03	7.99	8.17	**7.48**
MH_04	9.22	14.42	**9.14**
MH_05	10.82	13.22	**10.37**
Weighted Average	7.13	8.62	**6.69**

sequence of the Euroc MH dataset [42] and the time costs of the global pose-graph optimization. The compared variants were: 1) SP-VIDS: The standard pose-graph optimization was utilized to substitute our two-stage one; 2) LM-VIDS: The Levenberg-Marquadt method, a typical non-linear optimization scheme, rather than our EM-based smoothing was adopted in the second stage of our optimization pipeline. The RMSEs were summarized in Table II and the relationships between the time costs and the number of frames involved in the optimization were illustrated in Fig. 9. From the results, it can be seen that TES-CVIDS outperforms its other two variants in terms of both accuracy and speed, implying that our proposed two-stage optimization pipeline is crucial in guaranteeing the localization performance of TES-CVIDS.

TABLE III
REPROJECTED DEPTH ERRORS (CM) AND ERROR RATIOS OF COMPARED MAP REPRESENTATIONS ON DIFFERENT SCENES

Scenes	S1		S2		S3	
	Error	Ratio	Error	Ratio	Error	Ratio
Standard	5.77	2.49%	5.41	2.34%	7.35	3.38%
Ours	**5.49**	**2.37%**	**5.24**	**2.27%**	**6.68**	**3.07%**

Ablation Study for the Sub-map Representation: In the client-end of our TES-CVIDS, our proposed probabilistic TSDF representation was utilized. To verify the effectiveness of such a representation, some typical samples of sub-maps represented by both our proposed probabilistic TSDF form and the standard TSDF form were shown in Fig. 10. As shown in the figure, substituting the standard TSDF map with our proposed representation, the outliers in the sub-maps can be eliminated effectively. Besides, we also recorded the reprojected depth errors of these two compared map representations in each scene of our own collected dataset to evaluate the effectiveness of our representation quantitatively. The evaluation results were offered in Table III. It's worth mentioning that, to make the comparison more intuitive, we also provided the ratio of reprojected depth error to the average ground-truth depth. From Table III, it can be seen that the mapping accuracy of our TES-CVIDS can be obviously enhanced by integrating our proposed outlier-aware probabilistic TSDF representation.

VIII. CONCLUSION

In this paper, we studied a practical problem, collaborative localization and dense mapping for the multi-agent systems, and proposed a novel collaborative dense SLAM framework, namely TES-CVIDS. In TES-CVIDS, features and poses are packed and then sent to the central server from the client-end. Besides, RGBD images and corresponding poses are utilized to construct sub-maps, which are in the form of our proposed outlier-aware probabilistic TSDF representation. Instead of utilizing the most common image based transmission, the sub-maps are sent to the server and then be bound with corresponding key-frames. At the server-end, after aligning the local coordinate systems of different agents, the co-localization can be complete accurately and efficiently using our two-stage pose-graph optimization pipeline. Based on the accurate poses in a unified reference coordinate system of all key-frames, TES-CVIDS fuses the received sub-maps and reconstructs the scene densely. One eminent feature of TES-CVIDS is that, based on our proposed adaptive depth search mechanism, sub-maps can be recomposed to maintain the consistency between localization and mapping in the event of key-frame pose changes. The experimental results corroborate the superior performance of TES-CVIDS.

REFERENCES

[1] H. Liu, G. Zhang, and H. Bao, "Robust keyframe-based monocular SLAM for augmented reality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2016, pp. 1–10.

[2] D. Ramadasan, M. Chevaldonne, and T. Chateau, "Real-time SLAM for static multi-objects learning and tracking applied to augmented reality applications," in *Proc. IEEE Virtual Reality*, 2015, pp. 267–268.

[3] C. Papaioannidis, I. Mademlis, and I. Pitas, "Fast CNN-based single-person 2D human pose estimation for autonomous systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1262–1275, Mar. 2023.

[4] X. Shao, L. Zhang, T. Zhang, Y. Shen, and Y. Zhou, "MOFISSLAM: A multi-object semantic SLAM system with front-view, inertial and surround-view sensors for indoor parking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4788–4803, Jul. 2022.

[5] X. Yuwen, H. Zhang, F. Yan, and L. Chen, "Gaze control for active visual SLAM via panoramic cost map," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1813–1825, Feb. 2023.

[6] M. Eskandari and A. V. Savkin, "SLAPS: Simultaneous localization and phase shift for a RIS-equipped UAV in 5G/6G wireless communication networks," *IEEE Trans. Intell. Veh.*, vol. 8, no. 12, pp. 4722–4733, Dec. 2023.

[7] K. Nielsen and G. Hendeby, "Multi-hypothesis SLAM for non-static environments with reoccurring landmarks," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 3191–3203, Apr. 2023.

[8] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Field Robot.*, vol. 34, no. 3, pp. 314–334, 2015.

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

[10] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Automat. Lett.*, vol. 2, no. 2, pp. 796–803, Apr. 2017.

[11] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. IEEE Int. Symp. Mixed Augemented Reality*, 2011, pp. 127–136.

[12] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, Sep. 1987.

[13] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-D mapping with an RGB-D camera," *IEEE Trans. Robot.*, vol. 30, no. 1, pp. 177–187, Feb. 2014.

[14] A. Concha and J. Civera, "RGBDTAM: A cost-effective and accurate RGB-D tracking and mapping system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 6756–6763.

[15] G. Silveira, E. Malis, and P. Rives, "An efficient direct approach to visual SLAM," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 969–979, Oct. 2008.

[16] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–18, 2017.

[17] Q. Sun, J. Yuan, X. Zhang, and F. Duan, "Plane-Edge-SLAM: Seamless fusion of planes and edges for SLAM in indoor environments," *IEEE Trans. Automat. Sci. Eng.*, vol. 18, no. 4, pp. 2061–2075, Oct. 2021.

[18] T. Schops, T. Sattler, and M. Pollefeys, "Bad SLAM: Bundle adjusted direct RGB-D slam," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 134–144.

[19] Z. Shan, R. Li, and S. Schwertfeger, "RGBD-inertial trajectory estimation and mapping for ground robots," in *Sensors*, vol. 19, no. 10, pp. 2251–2279, 2019.

[20] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[21] V. Pradeep, C. Rhemann, S. Izadi, C. Zach, M. Bleyer, and S. Bathiche, "MonoFusion: Real-time 3D reconstruction of small scenes with a single web camera," in *Proc. IEEE/ACM Int. Symp. Mixed Augmented Reality*, 2013, pp. 83–88.

[22] M. Pizzoli, C. Forster, and D. Scaramuzza, "REMODE: Probabilistic, monocular dense reconstruction in real time," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 2609–2616.

[23] Z. Yang, F. Gao, and S. Shen, "Real-time monocular dense mapping on aerial robots using visual-inertial fusion," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 4552–4559.

[24] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3917–3925.

[25] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1983–1992.

[26] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9780–9790.

[27] D. Ren, J. Zheng, J. Cai, J. Li, and J. Zhang, "ExtrudeNet: Unsupervised inverse sketch-and-extrude for shape parsing," in *Proc. Eur. Conf. Comput. Vis*., 2022, pp. 482–498.
[28] C. Forster, S. Lynen, L. Kneip, and D. Scaramuzza, "Collaborative monocular SLAM with multiple micro aerial vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst*., 2013, pp. 3962–3970.
[29] D. Zou and P. Tan, "CoSLAM: Collaborative visual SLAM in dynamic environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 354–366, Feb. 2013.
[30] I. Deutsch, M. Liu, and R. Siegwart, "A framework for multi-robot pose graph SLAM," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot*., 2016, pp. 567–572.
[31] P. Schmuck and M. Chli, "CCM-SLAM: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams," *Int. J. Field Robot.*, vol. 36, no. 4, pp. 763–781, 2019.
[32] P. Y. Lajoie, B. Ramtoula, Y. Chang, L. Carlone, and G. Beltrame, "DOOR-SLAM: Distributed, online, and outlier resilient SLAM for robotic teams," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1656–1663, Apr. 2020.
[33] M. Karrer, P. Schmuck, and M. Chli, "CVI-SLAM: Collaborative visual-inertial SLAM," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 2762–2769, Oct. 2018.
[34] L. Riazuelo, J. Civera, and J. M. M. Montiel, "C^2TAM: A cloud framework for cooperative tracking and mapping," *Robot. Auton. Syst.*, vol. 62, no. 4, pp. 401–413, 2014.
[35] S. Golodetz, T. Cavallari, N. A. Lord, V. A. Prisacariu, D. W. Murray, and P. H. S. Torr, "Collaborative large-scale dense 3D reconstruction with online inter-agent pose optimisation," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 11, pp. 2895–2905, Nov. 2018.
[36] Y. Chang, Y. Tian, J. P. How, and L. Carlone, "Kimera-multi: A system for distributed multi-robot metric-semantic simultaneous localization and mapping," in *Proc. IEEE Int. Conf. Robot. Automat*., 2021, pp. 11210–11218.
[37] T. Zhang, L. Zhang, Y. Chen, and Y. Zhou, "CVIDS: A collaborative localization and dense mapping framework for multi-agent based visual-inertial SLAM," *IEEE Trans. Image Process.*, vol. 31, pp. 6562–6576, 2022.
[38] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image Vis. Comput.*, vol. 29, no. 7, pp. 434–441, 2011.
[39] I. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, 2011.
[40] J. Moré, "The Levenberg-Marquardt Algorithm: Implementation and Theory," in *Proc. Conf. Numer. Anal.*, 1978, pp. 105–116.
[41] X. Wu, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.
[42] M. Burri et al., "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.
[43] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.

Tianjun Zhang received the B.Sc. degree in 2019 from the School of Software Engineering, Tongji University, Shanghai, China, where he is currently working toward the Ph.D. degree. His research interests include collaborative SLAM, computer vision, and sensor calibration.

Lin Zhang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2006, respectively, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2011. From March 2011 to August 2011, he was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. In 2011, he joined the School of Software Engineering, Tongji University, Shanghai, where he is currently a Full Professor. His research interests include environment perception of intelligent vehicle, pattern recognition, computer vision, and perceptual image/video quality assessment. He is an Associate Editor for IEEE ROBOTICS AND AUTOMATION LETTERS, and *Journal of Visual Communication and Image Representation*. He was the recipient of the Young Scholar of Changjiang Scholars Program, Ministry of Education, China.

Fengyi Zhang received the B.Sc. degree from the School of Software Engineering, Shandong University, Jinan, China, in 2021. He is currently working toward the M.Sc. degree with the School of Software Engineering, Tongji University, Shanghai, China. His research interests include image enhancement, neural network compression, 3D reconstruction, and machine learning.

Shengjie Zhao (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the M.Sc. degree in electrical and computer engineering from China Aerospace Institute, Beijing, China, in 1991, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2004. He is currently the Dean of the College of Software Engineering and a Professor with the College of Software Engineering and the College of Electronics and Information Engineering, Tongji University, Shanghai, China. In previous postings, he conducted research with Lucent Technologies, Whippany, NJ, USA, and the China Aerospace Science and Industry Corporation, Beijing. He is a Fellow of the Thousand Talents Program of China and an Academician of the International Eurasian Academy of Sciences. His research interests include artificial intelligence, Big Data, wireless communications, image processing, and signal processing.

Yicong Zhou (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Hunan University, Changsha, China, and the M.Sc. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA. He is currently a Full Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include chaotic systems, multimedia security, computer vision, and machine learning. Dr. Zhou was the recipient of the Third Price of Macau Natural Science Award in 2014. He is an Associate Editor for *Neurocomputing*, *Journal of Visual Communication and Image Representation*, and *Signal Processing: Image Communication*. He is the Co-Chair of the Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He is a Senior Member of the International Society for Optical Engineering.