YOWO: You Only Walk Once to Jointly Map An Indoor Scene and Register Ceiling-mounted Cameras

Fan Yang, Sosuke Yamao, Ikuo Kusajima, Atsunori Moteki, Shoichi Masui, and Shan Jiang

***Abstract*—Using ceiling-mounted cameras (CMCs) for indoor visual capturing opens up a wide range of applications. However, registering CMCs to the target scene layout presents a challenging task. While manual registration with specialized tools is inefficient and costly, automatic registration with visual localization may yield poor results when visual ambiguity exists. To alleviate these issues, we propose a novel solution for jointly mapping an indoor scene and registering CMCs to the scene layout. Our approach involves equipping a mobile agent with a head-mounted RGB-D camera to traverse the entire scene once and synchronize CMCs to capture this mobile agent. The egocentric videos generate world-coordinate agent trajectories and the scene layout, while the videos of CMCs provide pseudo-scale agent trajectories and CMC relative poses. By correlating all the trajectories with their corresponding timestamps, the CMC relative poses can be aligned to the world-coordinate scene layout. Based on this initialization, a factor graph is customized to enable the joint optimization of ego-camera poses, scene layout, and CMC poses. We also develop a new dataset, setting the first benchmark for collaborative scene mapping and CMC registration[1] Experimental results indicate that our method not only effectively accomplishes two tasks within a unified framework, but also jointly enhances their performance. We thus provide a reliable tool to facilitate downstream position-aware applications.**

***Index Terms*—Multi-camera Video, Ego-centric Video, Indoor Scene Mapping, Camera Registration.**

I. Introduction

MULTI-CAMERA systems are commonly installed on ceilings in various indoor spaces, such as schools, retail stores, factories, and underground garages. Those scene-aware CMC captures have been employed for enhancing public safety surveillance, indoor virtual/augmented reality, and human-computer interaction [1]–[11]. To facilitate the functionality of scene-aware CMC capturing, a crucial aspect is mapping the scene layout and registering CMC six-degrees-of-freedom (6-DoF) poses to the scene layout.

In the field of scene mapping and CMC 6-DoF pose registration, two types of methodologies have been introduced. The first approach involves manual interventions: to map the scene layout using measuring tools (*e.g.*, rulers), and, to register CMC 6-DoF poses using calibration devices (*e.g.*, 3D calibration cubes) with established positions. Markers on the calibration devices are associated with CMCs to enable scene-aware 6-DoF pose registration [18]–[21]. However, this method is not scalable and proves to be inefficient. The second approach involves simultaneous localization and mapping (SLAM) and visual localization [22]–[30]. Following the 3D scene reconstruction via SLAM, it becomes feasible to align CMC 2D captures with the 3D scene features and subsequently derive correlated CMC 6-DoF poses (Fig. 1 (ii)). However, this approach could potentially be vulnerable to visual ambiguity of scene features, especially when considering the disparity between CMC captures and ego-camera captures (Fig. 2).

In our pursuit of automating the process and mitigating visual ambiguity, we look into another category of works—CMC *relative* 6-DoF pose estimation—that employs distinguishable mobile keypoints (*e.g.*, human skeletons) as cross-camera matching features [12], [13], [17], [31]–[38]. They propose that mobile keypoints could sustain robustness despite variations in viewpoint and viewing conditions across all pertinent views. Even though those approaches exhibit a commendable capability to mitigate the visual ambiguity inherent in scene static keypoints, their resulting CMC 6-DoF poses have not been automatically aligned to the scene layout, and, CMCs in isolated rooms cannot be correlated (Fig. 1 (i)).

Drawing inspiration from the above pioneering research, we propose a novel framework to jointly optimize mapping an indoor scene and registering CMC 6-DoF poses to this scene. Specifically, our methodology involves equipping a mobile agent, such as a robot or a person, with a head-mounted RGB-D camera to traverse the entire scene. Meanwhile, CMCs are synchronized to capture this mobile agent. While the ego-camera captures yield agent trajectories and the scene layout in the world coordinate, the CMC captures provide pseudo-scale agent trajectories and CMC relative 6-DoF poses. By correlating all the trajectories with their corresponding timestamps, the CMC relative 6-DoF poses can be aligned to the world-coordinate scene layout. Based on this initialization, a factor graph [39], [40] is tailored to enable the joint optimization of ego-camera 6-DoF poses, scene layout, and CMC 6-DoF poses (Fig. 1 (iii)). Once the joint optimization is complete, we obtain a world-coordinate scene layout and CMC 6-DoF poses registered to this layout, which enables a wide range of subsequent applications. Our solution is user-friendly. As implied by our title, consider yourself as a mobile agent, then "*you only walk once (YOWO)*" in the indoor scene to approach the results.

As depicted in Tab. I, our preferred settings markedly deviate from those of existing datasets, and therefore, we propose a novel dataset tailored to our specific requirements. We employ simulation engines, AI2THOR [41], [42] and Gym-UnrealCV [43], [44], to generate synthetic data for collaborative CMC and ego-camera capture. Leveraging our custom dataset, we conducted a comparative analysis of our approach against well-established methodologies in the field. The experimental results underscore the effectiveness and

Corresponding author: Fan Yang (fan.yang@fujitsu.com).
All authors are with Fujitsu Research, Japan.
[1]https://sites.google.com/view/yowo/home.

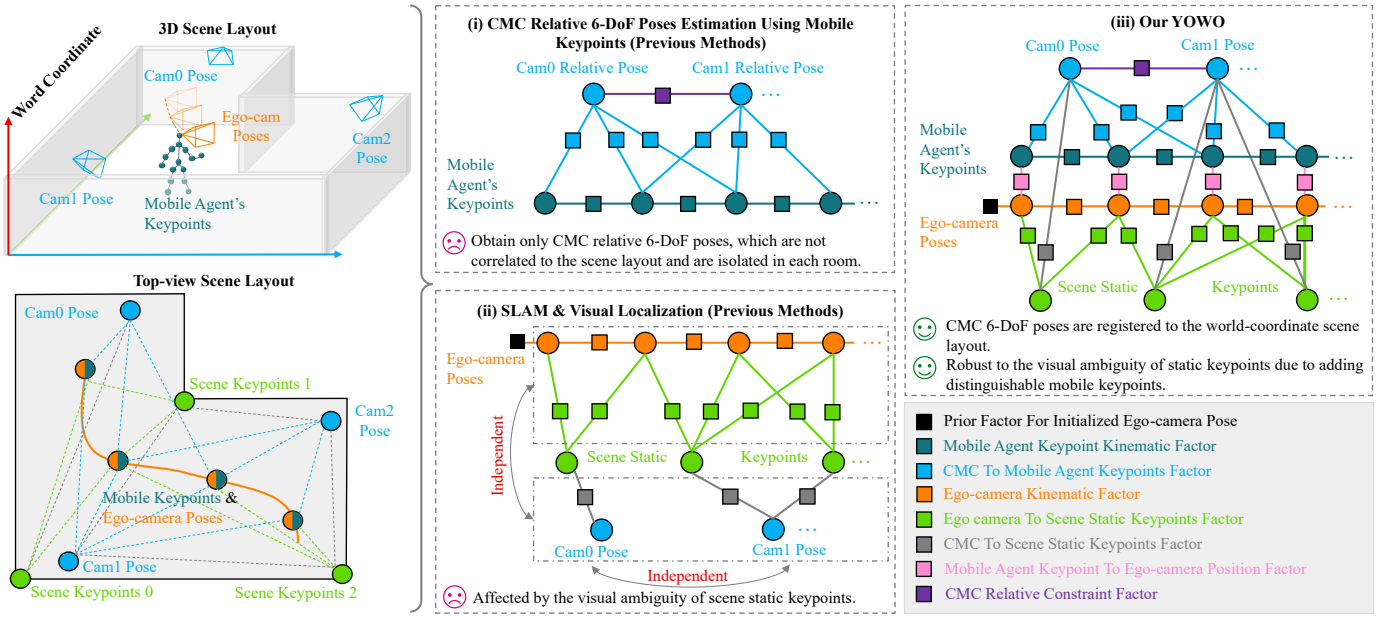

Fig. 1: **Overview of scene mapping and CMC 6-DoF pose registration**. While using mobile keypoints [12], [13] can estimate robust CMC relative 6-DoF poses (i), the application of SLAM & visual localization [14], [15] can generate CMC 6-DoF poses registered to the scene (ii). We bring their merits into our YOWO (iii).

TABLE I: **Dataset comparison.** We offer a novel dataset for collaborative CMC and ego-camera capturing.

Dataset	Target Task	Scene Type	No. Ego/Multi Cameras	Multi-camera Co-visibility	Captures	Matching Keypoints	Estimated Camera Vertical Height
OMECC [13]	Human Pose Estimation & CMC Relative 6-DoF Poses Estimation	Indoor	0, 20	Medium & High	Multi-view: RGB	Mobile	> 2 m
InLoc [14]	Scene Reconstruction & Visual Localization	Indoor	1, 0	-	Ego-view: RGB-D	Static	< 2 m
Egobody [16]	Human Pose Estimation	Indoor	1, 5	High	Multi-view: RGB-D Ego-view: RGB-D	Static	< 2 m
Egohumans [17]	Human Pose Estimation & Multi-human Tracking	In/Outdoor	4, 15	High	Multi-view: RGB Ego-view: RGB	Static	< 2 m
Ours	Map Scene Layout & Register CMC 6-DoF Poses To Scene	Indoor	1, 5-17	Low & Medium & High	Multi-view: RGB Ego-view: RGB-D	Static & Mobile	> 2 m

efficiency of our YOWO.

Our **contributions** are three-fold:

- We propose a versatile solution for jointly mapping an indoor scene and registering CMCs to the scene layout. Existing studies have not jointly optimized these two tasks, which limits their performance and effectiveness.
- We contribute the first CMC and ego-camera collaborative capturing dataset, for the joint evaluation of scene layout mapping and CMC 6-DoF pose registration.
- We perform a comprehensive evaluation of our method under diverse conditions, demonstrating that our approach can achieve robust and scalable performance. This capability facilitates a broad spectrum of downstream position-aware applications in indoor environments.

II. Related Work

A. SLAM And Visual Localization

By employing the power of SLAM or Structure-from-Motion (SfM) methods [22], [29], [45]–[50], the egocentric capture has been commonly applied for 3D indoor scene reconstruction [51]–[54]. Once the indoor scene has been reconstructed, visual localization can be applied to estimate the most probable camera 6-DoF pose for an arbitrary capture [14], [22], [23], [25]–[27], [30], [54]–[57]. However, static features extracted from the scene [58], [59], which are relied upon by most visual localization methods, could be significantly affected by ambiguous textures, symmetry structures, or even the changing structures within an indoor scene [14], [25], [26], [60]–[65]. As shown in Fig. 2, the disparity in perspective between the ego camera and CMCs further intensifies these challenges [25]. Furthermore, unlike the ego camera, our target CMCs remain static, and thus, we cannot leverage the global sequential camera poses to eliminate the camera localization errors [26]. Consequently, these issues undermine the performance of even the most accurate visual localization methods that follow a coarse-to-fine localization paradigm [15], [55], [56], [66]–[68].

Besides, although observations from the ego camera and

the CMCs could be strongly correlated under a synchronized configuration, extant literature has not yet leveraged this property. In our work, observations from the ego camera and CMCs are employed collaboratively to enhance both egocentric scene mapping and CMC registration. Specifically, instead of estimating independent CMC 6-DoF poses using conventional visual localization, observed mobile keypoints can be used together to form cross-view matching between CMCs and further infer their relative 6-DoF poses. Additionally, due to the lack of the global positioning signal, ego-camera SLAM may encounter position drift artifacts in indoor environments [22], [48], [54]. Our method leverages CMC observations to generate pseudo-scale ego-camera trajectories, thereby reducing ego-camera SLAM drift errors through fusion.

B. CMC Relative Pose Estimation Using Mobile Keypoints

Numerous previous studies [12], [12], [13], [31]–[38], [69], [70] have sought to automate the estimation of CMC relative pose by leveraging observed mobile keypoints. For instance, individuals observed across different cameras can be matched based on their appearance similarities, and the positions of these matched persons are used to estimate CMC relative 6-DoF poses [32], [69]. To improve the estimation performance, several investigations [33], [34], [38], [70] utilize multi-person body keypoints to achieve more accurate results. To circumvent the challenges in cross-camera multi-person matching, another group of research [12], [13], [31], [35], [37] execute collaborative CMC relative 6-DoF pose estimation on a single person.

Though these studies have significantly contributed to the advancement of CMC relative 6-DoF pose estimation, the capability to autonomously align CMC 6-DoF poses with the real scene layout has not been considered. In these works, scene mapping is treated as a separate task. However, this separation may limit potential applications. For example, in indoor surveillance, where map information is crucial, it is essential to register CMC 6-DoF poses to the scene layout. To remedy this limitation, our YOWO integrates the ego-camera scene mapping and CMC 6-DoF estimation into a unified framework to optimize them jointly.

C. Ego-camera And CMC Collaboration

Prior studies, as exemplified by [16], [17], [53], [71]–[74], have made notable progress in exploring the collaboration between ego cameras and CMCs for human/robot pose estimation and tracking. Whereas our YOWO shares some similar hardware setup to these studies, we diverge from their focus and expand our scope to address a novel challenge: the collaboration between scene mapping and CMC 6-DoF poses registration.

In terms of the approach, previous studies obtain CMC 6-DoF poses as an independent step before proceeding ego-camera and CMC collaboration [16], [17], [72]–[74]. For instance, although Egohumans [17] targets human pose estimation with the collaboration of ego camera and CMCs, it applies a classical visual localization approach to obtain CMC 6-DoF poses before its human pose estimation. In contrast, our YOWO incorporates mobile keypoints to address the limitations of classical visual localization. Additionally, unlike existing methods that exclusively utilize factor graphs to optimize either scene mapping or CMC relative poses, our YOWO forms a novel factor graph to jointly optimize both.

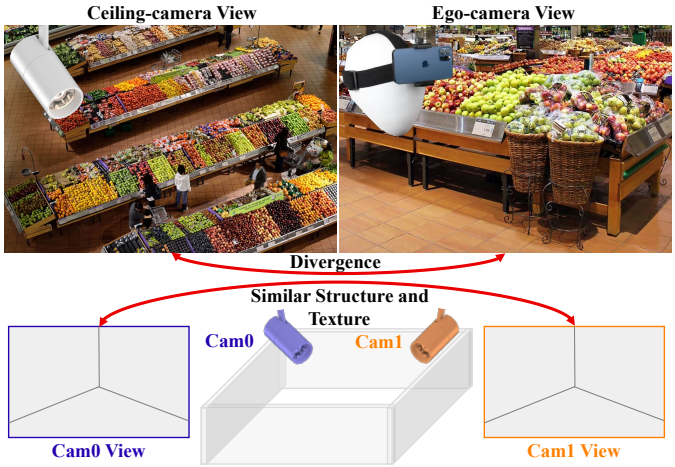

Fig. 2: **Visual ambiguity.** Top: the divergence between CMC and ego-camera views in a supermarket scene. Bottom: the similar captures between different CMCs.

In terms of the dataset, several studies [16], [17] do not fully capture scenes in their egocentric recording, which prevents the complete reconstruction of the scene layout. Some research [53], [73], [75] opt for multi-camera setups to obtain ground-truth (GT) agent trajectories but exclude these multi-camera recordings from their released datasets. Some works, *e.g.*, [76], capture top-view videos utilizing mobile drone cameras other than our static CMCs. Other studies, *e.g.*, [71], capture only egocentric RGB videos, which is insufficient to recover the real-scale scene layout. Therefore, we specifically create a novel dataset for our experimental analysis, as shown in Tab. I.

III. Approach

Problem formulation. Our study includes an RGB-D ego camera with the index c_e and multiple CMCs with indices $\{c_0, \cdots, c_n\}$. We suppose that all cameras have calibrated intrinsic parameters denoted as $\{\mathbf{K}_{c_e}, \mathbf{K}_{c_0}, \cdots, \mathbf{K}_{c_n}\} \in \mathbb{R}^{3\times3}$, which are provided by camera logs, or, by using calibration algorithms [18], [77]. Our study yields two primary outputs: (i) The scene point clouds comprise $N_{\mathcal{M}}$ points, denoted by $\{{}^w\mathbf{X}_0^{map}, \cdots, {}^w\mathbf{X}_{N_{\mathcal{M}}}^{map}\} \in \mathbb{R}^3$, which are further parameterized into 3D planes to represent the scene layout; (ii) The CMC 6-DoF poses registered to the scene layout, represented by matrices $\{[{}^w\mathbf{R}_{c_0}|{}^w\mathbf{X}_{c_0}], \cdots, [{}^w\mathbf{R}_{c_n}|{}^w\mathbf{X}_{c_n}]\} \in \mathbb{R}^{3\times4}$, where ${}^w\mathbf{R}_{c_i}$ and ${}^w\mathbf{X}_{c_i}$ are respectively the camera's orientation and position in the world coordinate. In this work, the superscripts w and l denote the coordinate systems for the world and the local, respectively.

Overview. YOWO consists of three primary processes: ego-camera processing, CMC processing, and collaborative processing (Fig. 3). In Sec. III-A, we introduce ego-camera processing, which includes using SLAM to acquire point

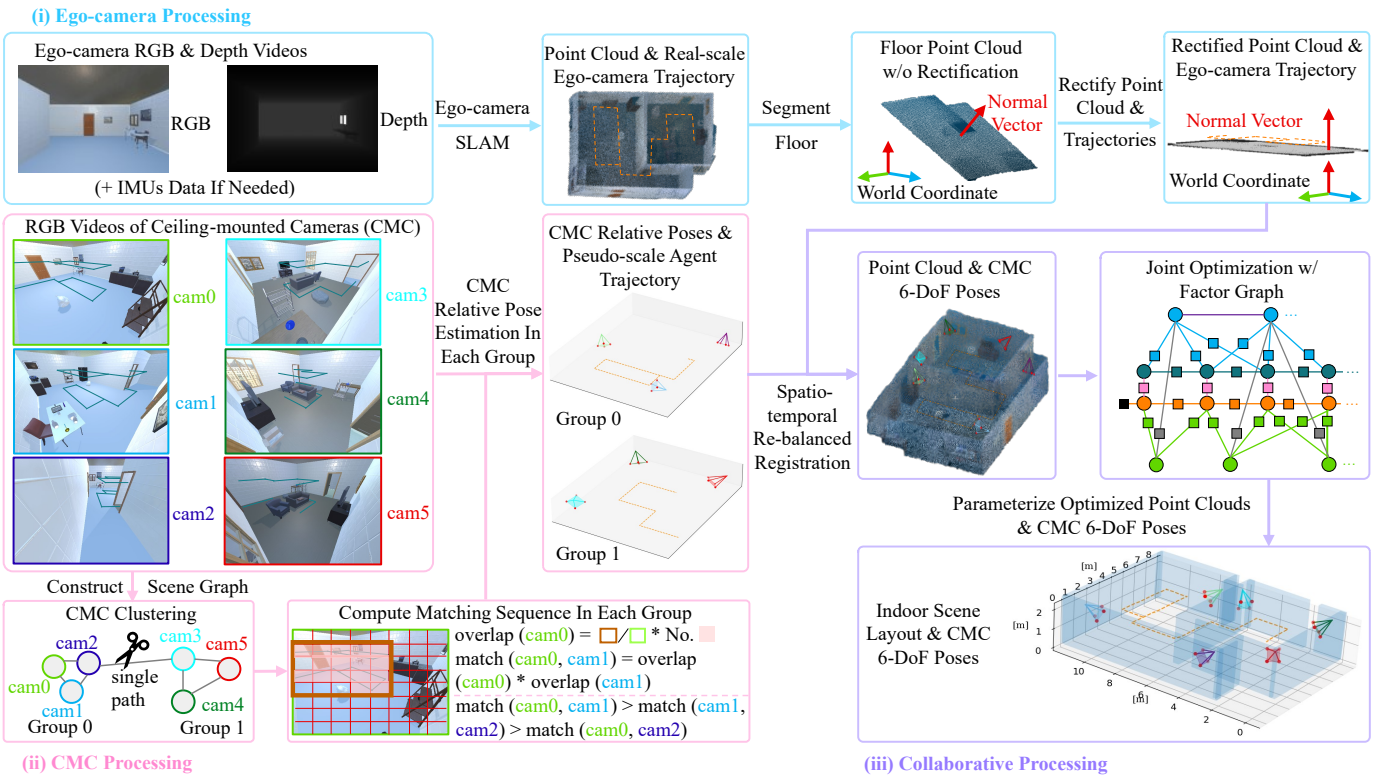

Fig. 3: **Architecture of YOWO**. YOWO mainly includes three processes: ego-camera processing (Sec. III-A), CMC processing (Sec. III-B), and collaborative processing (Sec. III-C). RGB/RGB-D videos, possibly with Inertial Measurement Unit (IMU) data, are the inputs, while the outputs are the scene layout and CMC 6-DoF poses registered to it.

clouds and ego-camera trajectories. We further elucidate the rectification process involved. We then detail CMC processing in Sec. III-B, including CMC scene graph clustering and CMC relative pose estimation within each group. In Sec. III-C, we introduce a novel spatiotemporal rebalanced registration algorithm for correlating ego-camera and CMC trajectories, which facilitates the alignment of CMC poses with point clouds. Based on this initialization, a factor graph is customized to enable the collaborative optimization of ego-camera poses, scene layout, and CMC poses. The optimized results are transformed into parameterized indoor scene layout and CMC 6-DoF poses. Given that our outputs are primarily designed for downstream applications, an offline approach is suitable.

A. Ego-camera Processing

We start by equipping a mobile agent with an ego camera positioned at the head, and let the agent traverse the entire scene to capture RGB-D images, denoted by $\left\{\left\langle \mathbf{I}^{rgb}_{c_e,f}, \mathbf{I}^{d}_{c_e,f} \right\rangle\right\}_{\forall f \in \mathcal{F}}$, where $\mathcal{F}$ is the set of frames. Using $\left\{\mathbf{I}^{d}_{c_e,f}\right\}_{\forall f \in \mathcal{F}}$, we initialize ego-camera motions using Iterative Closest Point (ICP) [78], [79] with multiway registration [80]. This approach circumvents the effect of visual noise and ambiguities in our initialization. Meanwhile, inertial measurement units (IMUs) attached to the ego camera could be employed to improve the results [26], [48]. By setting the initial world coordinate at the first ego-camera pose, the ego-camera motions are integrated into the ego-camera 6-DoF poses, as denoted by $\{[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]\}_{\forall f \in \mathcal{F}}$. Next, $\left\{\mathbf{I}^{rgb}_{c_e,f}\right\}_{\forall f \in \mathcal{F}}$ are incorporated to refine the ego-camera 6-DoF poses. We apply keypoint detection and matching algorithms [58], [59], [81] to obtain matched 2D keypoints from consecutive frames. To address challenging indoor appearance, in addition to directly detecting 2D scene keypoints, we also apply a line detection and matching method [82] to obtain matched line endpoints. To this end, 2D scene keypoints, captured by the ego camera, are represented by $\{\mathbf{x}_{c_e,f,m}\}_{\forall m \in \mathcal{M}, \forall f \in \mathcal{F}}$, where $\mathcal{M}$ denotes the set of scene keypoint indices. In addition to using the Epipolar constraint [83], the initialized ego-camera poses and depth information are employed to filter out inconsistent keypoints [84]. Specifically, based on preset ego-camera poses, we separately estimate 3D points by using paired 2D keypoint triangulation and depth-to-3D mapping [85]. If the Euclidean distance between identical 3D points exceeds a certain threshold, we reject the corresponding 2D keypoints [85]. Filtered 2D keypoints and corresponding 3D points form 2D-3D matches $\{\langle \mathbf{x}_{c_e,f,m}, {}^{w}\mathbf{X}^{map}_{m} \rangle\}_{\forall m \in \mathcal{M}, \forall f \in \mathcal{F}}$. For the sake of simplicity, we employ a basic method to initialize scene reconstruction in YOWO. However, it is flexible to adopt more advanced SfM/SLAM techniques to construct $\{\langle \mathbf{x}_{c_e,f,m}, {}^{w}\mathbf{X}^{map}_{m} \rangle\}_{\forall m \in \mathcal{M}, \forall f \in \mathcal{F}}$ and $\{[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]\}_{\forall f \in \mathcal{F}}$.

We apply a bundle adjustment (BA) [85], [86] to jointly adjust ego-camera poses and 2D-3D matches, which is for-

mulated by

$$\underset{\substack{\{[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]\},\\ \{{}^{w}\mathbf{X}_{m}^{map}\}}}{\arg\min} \sum_{\forall f\in\mathcal{F}}\sum_{\forall m\in\mathcal{M}} \left\|\varepsilon_{ego}\left([{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}],\ {}^{w}\mathbf{X}_{m}^{map}\right)\right\|^2, \tag{1}$$

where the error to be minimized are represented by

$$\begin{aligned}&\left\|\varepsilon_{ego}\left([{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}],\ {}^{w}\mathbf{X}_{m}^{map}\right)\right\|^2\\ =&v_{f,m}\ \rho\left\|\mathbf{x}_{c_e,f,m}-\pi\left(\mathbf{K}_{c_e}\left[{}^{w}\mathbf{R}_{c_e,f}^{\top}|-{}^{w}\mathbf{R}_{c_e,f}^{\top}\,{}^{w}\mathbf{X}_{c_e,f}\right],\right.\right.\\ &\left.\left.{}^{w}\mathbf{X}_{m}^{map}\right)\right\|^2,\end{aligned} \tag{2}$$

where ρ is a Huber robust M-estimator, $\pi(\cdot)$ represents the projection function that maps 3D points to 2D points given the camera pose. The binary variable $v_{f,m}$ equals to 1 if ${}^{w}\mathbf{X}_{m}^{map}$ is visible at frame f, and 0 otherwise. The formula is optimized by employing the Levenberg–Marquardt algorithm [87]. We construct the point clouds using the optimized ego-camera poses and RGB-D images.

In practice, it is challenging to perfectly initialize the ego-camera pose such that its vertical direction is perpendicular to the ground plane. The alignment bias is subsequently inherited by the constructed point clouds (Fig. 3). To accurately map the scene layout, we compute a rotation matrix $\mathbf{R}_{rect}$ to rectify the point clouds in the vertical direction. We apply planar patch detection [88] to detect planes on point clouds. The plane with the most points is considered as the ground plane. Let $\mathbf{v}_{gp}$ represent the normal vector of the ground plane, $\mathbf{v}_w$ denote the initialized world-coordinate vertical vector, and $\mathbb{I}_{3\times3}$ symbolize an identity matrix, the rotation matrix is given by

$$\begin{aligned}&\mathbf{R}_{rect}=\mathbb{I}_{3\times3}+[\mathbf{g}]_{\times}+[\mathbf{g}]_{\times}^{2}\frac{1-\mathbf{v}_{gp}\cdot\mathbf{v}_{w}}{\|\mathbf{g}\|^{2}},\\ &\text{with } \mathbf{g}=\mathbf{v}_{gp}\times\mathbf{v}_{w} \text{ and } [\mathbf{g}]_{\times}\overset{\text{def}}{=}\begin{bmatrix}0&-\mathbf{g}_3&\mathbf{g}_2\\ \mathbf{g}_3&0&-\mathbf{g}_1\\ -\mathbf{g}_2&\mathbf{g}_1&0\end{bmatrix}.\end{aligned} \tag{3}$$

We then use $\mathbf{R}_{rect}$ to rectify $\{{}^{w}\mathbf{X}_{m}^{map}\}$ and $\{[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]\}$ in the vertical direction:

$${}^{w}\mathbf{X}_{m}^{map}=\mathbf{R}_{rect}\,{}^{w}\mathbf{X}_{m}^{map} \tag{4}$$

$$[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]=\mathbf{R}_{rect}\,[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]. \tag{5}$$

B. CMC Processing

Our mobile agent can either be a human or a robot. To estimate the 2D keypoint sequence of the agent for each CMC, we employ off-the-shelf 2D keypoint estimation and tracking algorithms [89]–[92] for a human agent. For a robot agent, we initially annotate samples for its top and bottom points, subsequently fine-tuning the same algorithms to estimate its keypoints. We skip the details of annotation and training here. To eliminate jitters in the mobile keypoint sequence, we further apply a Savitzky–Golay filter [93] to smooth the data.

Given our assumption that there is only one mobile agent in the scene, we can establish cross-camera correspondences by referencing the timestamps and keypoint indices. For instance, at the frame f, if a mobile keypoint with index o is concurrently observed by a set of ceiling-mounted cameras $\mathcal{C}^{cmc}$, the 2D keypoint set $\{\mathbf{x}_{c_i,f,o}\}_{\forall c_i\in\mathcal{C}^{cmc}}$ corresponds to the same 3D keypoint $\mathbf{X}_{a,f,o}^{cmc}$. Thus, although we follow incremental SfM [45], [46] to estimate CMC relative 6-DoF poses, we can skip the steps of image retrieval [94], [95] and feature matching [59], [81], [96]. However, due to the physical barriers posed by indoor walls, CMCs may be divided into separate spaces, and therefore, we cluster them into groups based on their co-observations [15], [66], [97]. Given the unique nature of mobile keypoints, we let $\mathcal{A}$ and $\mathcal{B}$ respectively denote the sets of co-visible mobile keypoints of camera c_i and c_j, as determined by their timestamps and keypoint indices. We then exclude noisy samples by estimating the fundamental matrix ${}^{c_j}\mathbf{F}_{c_i}$ and verifying the Epipolar constraint error with a MAGSAC scheme [98] (*i.e.*, a threshold-unsensitive RANSAC [99]). We formulate the inlier counting for $\mathcal{A}$ and $\mathcal{B}$ as

$$\begin{aligned}&\text{inliers}(c_i,c_j)=\sum_{\forall\mathbf{x}_{c_i,f,o}\in\mathcal{A}}\ \sum_{\forall\mathbf{x}_{c_j,f,o}\in\mathcal{B}}\\ &\mathbb{1}\left[\left\|\mathbf{x}_{c_i,f,o}^{\top}\,{}^{c_j}\mathbf{F}_{c_i}\mathbf{x}_{c_j,f,o}\right\|_2\le\text{RANSAC Threshold}\right].\end{aligned} \tag{6}$$

Using Eq. 6, we construct a scene graph [46], [100] and perform graph clustering [101] to divide CMCs into groups. Within the same group, we impose two conditions between any two elements: (i) the presence of at least eight matching pairs [85], [102], (ii) a minimum of two paths connecting one element to another, to ensure the 2D-3D matches are available for any camera. We apply these criteria to guarantee that stable CMC relative poses can be formed.

Within a camera group, we assume that the inlier samples of $\mathcal{A}$ and $\mathcal{B}$ form new sets $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$. Inspired by [26], [46], [103], [104], we define a binary grid matrix $\mathbf{G}_{c_i}$ to quantize $\tilde{\mathcal{A}}$ and formulate our spatial overlap score as

$$\mathbf{G}_{c_i}[r,l]=\begin{cases}1, & \text{if } \exists\ \mathbf{x}_{c_i,f,o}\in[r,r+1)\times[l,l+1),\\ & \text{and } \forall\mathbf{x}_{c_i,f,o}\in\tilde{\mathcal{A}},\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

$$\begin{aligned}\text{overlap}(c_i)=&\frac{\left(\max_{r}\mathbf{G}_{c_i}-\min_{r}\mathbf{G}_{c_i}\right)\cdot\left(\max_{l}\mathbf{G}_{c_i}-\min_{l}\mathbf{G}_{c_i}\right)}{N_{GW}\cdot N_{GH}}\\ &\cdot\sum_{r\in N_{GH},l\in N_{GW}}\mathbf{G}_{c_i}[r,l],\end{aligned} \tag{8}$$

where $r\in N_{GH}$ and $l\in N_{GW}$ denote the row and column of $\mathbf{G}_{c_i}$, respectively, and $[r,r+1)\times[l,l+1)$ represents the grid cell. A higher overlap score generally results in a more accurate 6-DoF pose estimation. We obtain the overlap(c_j) for $\tilde{\mathcal{B}}$ with the same formula.

Since CMCs are typically installed with well-structured triangulation baselines and angles, we leave these factors aside. Thus, we select the initial pair of cameras with the largest matching score, computed using

$$\text{match}(c_i,c_j)=\text{overlap}(c_i)\cdot\text{overlap}(c_j). \tag{9}$$

Assuming the camera indices of the initial pair are c_i and

c_j, we compute their essential matrix ${}^{c_j}\mathbf{E}_{c_i}$ from previously obtained ${}^{c_j}\mathbf{F}_{c_i}$, as

$$ {}^{c_j}\mathbf{E}_{c_i} = \mathbf{K}_{c_i}^{\top}\, {}^{c_j}\mathbf{F}_{c_i}\mathbf{K}_{c_j}. \tag{10} $$

We then extract the rotation matrix ${}^{c_j}\mathbf{R}_{c_i}$ and the translation vector ${}^{c_j}\mathbf{t}_{c_i}$ by applying the singular value decomposition (SVD) to ${}^{c_j}\mathbf{E}_{c_i}$ [83], [85]. Among multiple solutions, we employ geometric verification to select the most suitable one [105]. The center of the reference camera c_i is selected as the local coordinate center for its group, as ${}^{l}\mathbf{R}_{c_i} = [\mathbb{I}_{3\times3}]$ and ${}^{l}\mathbf{X}_{c_i} = [0,0,0]$. The relative 6-DoF of camera c_j is $[{}^{l}\mathbf{R}_{c_j}|{}^{l}\mathbf{X}_{c_j}]$, with ${}^{l}\mathbf{R}_{c_j} = {}^{c_j}\mathbf{R}_{c_i}^{\top}$ and ${}^{l}\mathbf{X}_{c_j} = -{}^{c_j}\mathbf{R}_{c_i}^{\top}\, {}^{c_j}\mathbf{t}_{c_i}$. We further construct their projection matrices, as

$$ \mathbf{P}_{c_i} = \mathbf{K}_{c_i}[{}^{l}\mathbf{R}_{c_i}^{\top}| - {}^{l}\mathbf{R}_{c_i}^{\top}\, {}^{l}\mathbf{X}_{c_i}], \tag{11} $$

$$ \mathbf{P}_{c_j} = \mathbf{K}_{c_j}[{}^{l}\mathbf{R}_{c_j}^{\top}| - {}^{l}\mathbf{R}_{c_j}^{\top}\, {}^{l}\mathbf{X}_{c_j}], \tag{12} $$

and apply 3D triangulation [46], [85] to generate the local-coordinate 3D mobile keypoints $\left\{{}^{l}\mathbf{X}_{a,f,o}^{cmc}|\forall f \in \mathcal{F}, \forall o \in \mathcal{O}\right\}$, where $\mathcal{O}$ is the set of mobile keypoint indices.

Leveraging the timestamps and keypoint indices, we can establish 2D-3D matches between the estimated 3D points and the 2D points from a newly introduced camera within the same group. These matches can then be utilized to estimate the relative 6-DoF pose of the new camera, via PnP [106], [107] and RANSAC. Among multiple unmatched cameras, we prefer to initially match the camera that has the largest overlapping score between its 2D points to the existing 3D points. As such, we utilize Eqs. 7 and 8 to compute overlapping scores for all unmatched cameras within a group, and select the camera with the highest overlapping score to estimate its 6-DoF pose. Upon obtaining the new 6-DoF pose, we generate additional 3D mobile keypoints by integrating the new camera into the system. The overlapping scores should be updated after adding new 3D points. Repeating this process yields the full CMC relative 6-DoF poses and pseudo-scale agent trajectories within each group.

After obtaining CMC relative 6-DoF poses, we perform a BA to refine them within each group. In our CMC BA, we denote the error function for the 2D-3D projection constraint as $\varepsilon_{cmc}(\cdot)$. We minimize it to jointly optimize CMC relative 6-DoF poses $\left\{\left[{}^{l}\mathbf{R}_{c_i}|{}^{l}\mathbf{X}_{c_i}\right]\right\}$ and 3D mobile keypoints $\left\{{}^{l}\mathbf{X}_{a,f,o}^{cmc}\right\}$ in their pseudo-scale local coordinates. The formulation is as follows:

$$ \underset{\left\{\left[{}^{l}\mathbf{R}_{c_i}|{}^{l}\mathbf{X}_{c_i}\right]\right\},\left\{{}^{l}\mathbf{X}_{a,f,o}^{cmc}\right\}}{\arg\min} \sum_{\forall c_i \in \mathcal{C}^{cmc}} \sum_{\forall f \in \mathcal{F}} \sum_{\forall o \in \mathcal{O}} \left\| \varepsilon_{cmc}\left(\left[{}^{l}\mathbf{R}_{c_i}|{}^{l}\mathbf{X}_{c_i}\right], {}^{l}\mathbf{X}_{a,f,o}^{cmc}\right)\right\|^2, $$

where

$$ \begin{aligned} &\left\| \varepsilon_{cmc}\left(\left[{}^{l}\mathbf{R}_{c_i}|{}^{l}\mathbf{X}_{c_i}\right], {}^{l}\mathbf{X}_{a,f,o}^{cmc}\right)\right\|^2 \\ &= v_{c_i,f,o}\ \rho\left\|\mathbf{x}_{c_i,f,o} - \pi\left(\mathbf{K}_{c_i}\left[{}^{l}\mathbf{R}_{c_i}^{\top}|-{}^{l}\mathbf{R}_{c_i}^{\top}\, {}^{l}\mathbf{X}_{c_i}\right], {}^{l}\mathbf{X}_{a,f,o}^{cmc}\right)\right\|^2. \end{aligned} \tag{13} $$

Within a group, the set of cameras, frames, and keypoints are denoted by $\mathcal{C}^{cmc}$, $\mathcal{F}$, and $\mathcal{O}$. The binary variable $v_{c_i,f,o}$ represents the visibility of 3D-2D projection.

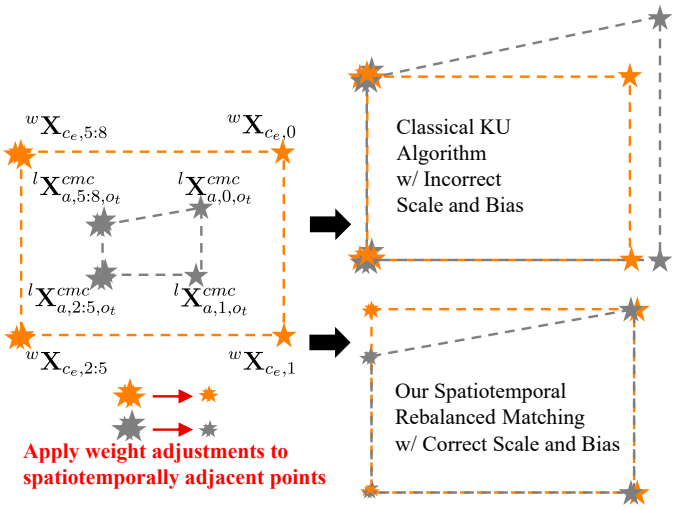

Fig. 4: **Comparison of classical KU matching and our spatiotemporal rebalanced matching.**

C. Collaborative Processing

Register CMCs to the scene. Using multiple mobile keypoints provides sufficient observations to apply geometric verifications for a robust CMC relative 6-DoF pose estimation. However, in our ego-camera processing, we only can obtain the ego-camera trajectory $\left\{{}^{w}\mathbf{X}_{c_e,f} \in \mathbb{R}^3\right\}$ for one position, which is closed to the agent's top keypoint in our setting. Therefore, we employ the agent's top keypoint trajectory, as denoted by $\left\{{}^{l}\mathbf{X}_{a,f,o_t}^{cmc} \in \mathbb{R}^3\right\}$, to approximate the local-coordinate ego-camera trajectory. We then align $\left\{{}^{l}\mathbf{X}_{a,f,o_t}^{cmc}\right\}$ with $\{{}^{w}\mathbf{X}_{c_e,f}\}$ to transform the CMC processing results to the world coordinate. Although there may exist a tiny discrepancy between the ego-camera position and the agent's top keypoint, this discrepancy becomes insignificant when CMCs are sufficiently distant from the agent. Furthermore, we will incorporate this discrepancy into our collaborative factor graph optimization with an appropriate uncertainty to tolerate it (see Eqs. 30 and 32).

In previous studies [13], [108], the Kabsch-Umeyama (KU) algorithm [109], [110] has been utilized to align relative CMC positions to the GT for evaluation purposes, or to realign relative CMC positions for iterative pose refinement. In our YOWO, the KU algorithm could be used for matching $\left\{{}^{l}\mathbf{X}_{a,f,o_t}^{cmc}\right\}$ to $\{{}^{w}\mathbf{X}_{c_e,f}\}$. We want to find the optimal rescaling factor ${}^{l}s_w$, rotation matrix ${}^{l}\mathbf{R}_w$, and translation vector ${}^{l}\mathbf{t}_w$ such that:

$$ \underset{{}^{l}s_w,{}^{l}\mathbf{R}_w,{}^{l}\mathbf{t}_w}{\arg\min} \sum_{\forall f \in \mathcal{F}} \left\| {}^{w}\mathbf{X}_{c_e,f} - \left({}^{l}s_w\, {}^{l}\mathbf{R}_w\, {}^{l}\mathbf{X}_{a,f,o_t}^{cmc} + {}^{l}\mathbf{t}_w\right)\right\|^2. \tag{14} $$

However, using Eq. 14 in YOWO presents a unique challenge: if the agent remains stationary for an extended period while capturing different views, the reference to timestamps will yield a large number of 3D points at the same position. This results in an overemphasis on this position, leading to biased matching, as depicted in Fig. 4.

To mitigate this issue, we recast the KU algorithm to a spa-

tiotemporal rebalanced matching algorithm. First, we adapt the single-linkage clustering [111] to allocate $\{{}^{w}\mathbf{X}_{c_e,f}\}$ into clusters $\{\phi_0, \phi_1, \cdots, \phi_{N_{clus}-1}\}$. This adaptation incorporates two conditions into the single-linkage clustering process, ensuring that any two arbitrary elements, $\{{}^{w}\mathbf{X}_{c_e,f_i}, {}^{w}\mathbf{X}_{c_e,f_j}\} \in \phi_k$, fulfill the following criteria:

$$\begin{cases} \left\| {}^{w}\mathbf{X}_{c_e,f_i} - {}^{w}\mathbf{X}_{c_e,f_j} \right\|_2 \leq \text{spatialThreshold} \\ |f_i - f_j| \leq \text{temporalThreshold} \end{cases} \tag{15}$$

Then, by using the clustering results, we add the rebalanced weight α_f to Eq. 14 and construct our rebalanced matching algorithm as

$$\begin{aligned} &\underset{{}^{l}s_w, {}^{l}\mathbf{R}_w, {}^{l}\mathbf{t}_w}{\arg\min} \sum_{\forall f \in \mathcal{F}} \alpha_f \left\| {}^{w}\mathbf{X}_{c_e,f} - ({}^{l}s_w \, {}^{l}\mathbf{R}_w \, {}^{l}\mathbf{X}^{cmc}_{a,f,o_t} + {}^{l}\mathbf{t}_w) \right\|^2, \\ &\text{where } \alpha_f = \frac{1}{N_{clus} N_{\phi_k}} \text{ if } {}^{w}\mathbf{X}_{c_e,f} \in \phi_k, \\ &\sum_{\forall f \in \mathcal{F}} \alpha_f = 1, \end{aligned} \tag{16}$$

where $\mathcal{F}$ denotes the co-visible frames between the ego camera and the CMCs, N_{clus} and N_{ϕ_k} represent the number of spatiotemporal clusters and elements of cluster ϕ_k, respectively.

The corresponding solution slightly deviates from the original KU algorithm and can be derived using the following steps. We begin by calculating the weighted centroids with $\{\alpha_f\}_{\forall f \in \mathcal{F}}$:

$${}^{l}\bar{\mathbf{X}}^{cmc}_{a,o_t} = \sum_{\forall f \in \mathcal{F}} \alpha_f \, {}^{l}\mathbf{X}^{cmc}_{a,f,o_t}, \quad {}^{w}\bar{\mathbf{X}}_{c_e} = \sum_{\forall f \in \mathcal{F}} \alpha_f \, {}^{w}\mathbf{X}_{c_e,f}. \tag{17}$$

Next, we compute the weighted cross-covariance matrix between matching sequences $\{{}^{w}\mathbf{X}_{c_e,f}\}_{\forall f \in \mathcal{F}}$ and $\left\{{}^{l}\mathbf{X}^{cmc}_{a,f,o_t}\right\}_{\forall f \in \mathcal{F}}$:

$$\begin{aligned} \mathbf{H} = &\begin{bmatrix} {}^{w}\mathbf{X}_{c_e,0} - {}^{w}\bar{\mathbf{X}}_{c_e} \\ {}^{w}\mathbf{X}_{c_e,1} - {}^{w}\bar{\mathbf{X}}_{c_e} \\ \vdots \\ {}^{w}\mathbf{X}_{c_e,N_{\mathcal{F}}-1} - {}^{w}\bar{\mathbf{X}}_{c_e} \end{bmatrix}^{\intercal} \operatorname{diag}\left(\alpha_0, \alpha_1, \cdots, \alpha_{N_{\mathcal{F}}-1}\right) \\ &\begin{bmatrix} {}^{l}\mathbf{X}^{cmc}_{a,0,o_t} - {}^{l}\bar{\mathbf{X}}^{cmc}_{a,o_t} \\ {}^{l}\mathbf{X}^{cmc}_{a,1,o_t} - {}^{l}\bar{\mathbf{X}}^{cmc}_{a,o_t} \\ \vdots \\ {}^{l}\mathbf{X}^{cmc}_{a,N_{\mathcal{F}}-1,o_t} - {}^{l}\bar{\mathbf{X}}^{cmc}_{a,o_t} \end{bmatrix}, \end{aligned} \tag{18}$$

where $N_{\mathcal{F}}$ represent the number of the set $\mathcal{F}$. We then compute the weighted covariance of sequences $\{{}^{w}\mathbf{X}_{c_e,f}\}_{\forall f \in \mathcal{F}}$:

$$\sigma^2_{{}^{w}\mathbf{X}_{c_e}} = \sum_{\forall f \in \mathcal{F}} \alpha_f \left\| {}^{w}\mathbf{X}_{c_e,f} - {}^{w}\bar{\mathbf{X}}_{c_e} \right\|^2 \tag{19}$$

The remaining calculations are identical to the original Kabsch-Umeyama algorithm:

$$\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^{\intercal}, \; d = \operatorname{sign}\big(\det\left(\mathbf{V}\mathbf{U}^{\intercal}\right)\big), \; \mathbf{S} = \operatorname{diag}\left(\underbrace{1, \cdots, 1}_{N_{\mathcal{F}}-1}, d\right) \tag{20}$$

The optimal rotation matrix ${}^{l}\mathbf{R}_w$ and scale factor ${}^{l}s_w$ are computed using the SVD of the matrix $\mathbf{H}$:

$${}^{l}\mathbf{R}_w = \mathbf{U}\mathbf{S}\mathbf{V}^{\intercal} \tag{21}$$

$${}^{l}s_w = \frac{\sigma^2_{{}^{w}\mathbf{X}_{c_e}}}{\operatorname{tr}(\Sigma\mathbf{S})} \tag{22}$$

Finally, the translation vector ${}^{l}\mathbf{t}_w$ is computed as follows:

$${}^{l}\mathbf{t}_w = {}^{w}\bar{\mathbf{X}}_{c_e} - {}^{l}s_w \, {}^{l}\mathbf{R}_w \, {}^{l}\bar{\mathbf{X}}^{cmc}_{a,o_t} \tag{23}$$

After that, we use the estimated $\left\{{}^{l}s_w, {}^{l}\mathbf{R}_w, {}^{l}\mathbf{t}_w\right\}$ to transform $\left\{{}^{l}\mathbf{X}^{cmc}_{a,f,o}\right\}$ and $\left\{\left[{}^{l}\mathbf{R}_{c_i} \middle| {}^{l}\mathbf{X}_{c_i}\right]\right\}$ from the local coordinate to the world coordinate:

$${}^{w}\mathbf{X}^{cmc}_{a,f,o} = {}^{l}s_w \, {}^{l}\mathbf{R}_w \, {}^{l}\mathbf{X}^{cmc}_{a,f,o} + {}^{l}\mathbf{t}_w \tag{24}$$

$${}^{w}\mathbf{R}_{c_i} = {}^{l}\mathbf{R}_w \, {}^{l}\mathbf{R}_{c_i} \tag{25}$$

$${}^{w}\mathbf{X}_{c_i} = {}^{l}s_w \, {}^{l}\mathbf{R}_w \, {}^{l}\mathbf{X}_{c_i} + {}^{l}\mathbf{t}_w. \tag{26}$$

We apply RGB-image feature detectors [58], [59] again to obtain 2D scene keypoints and their corresponding features for CMC RGB images. Then, we integrate CMC 6-DoF pose priors (*i.e.*, $\{[{}^{w}\mathbf{R}_{c_i}|{}^{w}\mathbf{X}_{c_i}]\}$) to match these 2D keypoints to 3D scene keypoints [15], which are estimated from our ego-camera processing. All 2D-3D matches, from both scene and mobile keypoints, are merged into a set of 2D-3D pairs represented by

$$\begin{aligned} \{\langle \mathbf{x}_{c_i,f,m}, {}^{w}\mathbf{X}_m \rangle \; | \; &\forall c_i \in \mathcal{C}^{cmc} \cup \mathcal{F}, \forall f \in \mathcal{F}, \\ &\forall m \in \mathcal{M} \cup \{\mathcal{F} \times \mathcal{O}\}\}. \end{aligned} \tag{27}$$

Joint optimization with factor graph. Within the same coordinate system, Eqs. 1 and 13 can be consolidated into a single equation. We use $\varepsilon_{\text{c2k}}(\cdot)$ to denote the 2D-3D projection constraints that encompass all camera poses and keypoints. It incorporates a joint optimization of the factors between: CMC poses and mobile-agent keypoints, CMC poses and scene keypoints, ego-camera poses and scene keypoints, and relative constraints of the CMC poses. This is visually represented in Fig. 1 and mathematically expressed in the subsequent equation:

$$\underset{\mathcal{T}, \mathcal{X}}{\arg\min} \sum_{\forall c_i \in \mathcal{C}^{cmc} \cup \mathcal{F}} \sum_{\forall f \in \mathcal{F}} \sum_{\forall m \in \mathcal{M} \cup \{\mathcal{F} \times \mathcal{O}\}} \varepsilon_{\text{c2k}}\left(\mathcal{T}, \mathcal{X}\right), \tag{28}$$

where all 3D keypoints and camera poses are merged into the set $\mathcal{X}$ and the set $\mathcal{T}$, respectively. In detail, we represent them by

$$\begin{aligned} \varepsilon_{\text{c2k}}\left(\mathcal{T}, \mathcal{X}\right) = &v_{c_i,f,m} \left\| \mathbf{x}_{c_i,f,m} - \pi\left(\mathbf{K}_{c_i}\left[{}^{w}\mathbf{R}^{\top}_{c_i} \middle| -{}^{w}\mathbf{R}^{\top}_{c_i} {}^{w}\mathbf{X}_{c_i}\right], \right.\right. \\ &\left.\left. {}^{w}\mathbf{X}_m\right)\right\|_2, \\ \mathcal{T} = &\{[{}^{w}\mathbf{R}_{c_i}|{}^{w}\mathbf{X}_{c_i}]\} \cup \{[{}^{w}\mathbf{R}_{c_e,f}|{}^{w}\mathbf{X}_{c_e,f}]\} \\ = &\{[{}^{w}\mathbf{R}_{c_i}| \; {}^{w}\mathbf{X}_{c_i}] | \forall c_i \in \mathcal{C}^{cmc} \cup \mathcal{F}\}, \\ \mathcal{X} = &\{{}^{w}\mathbf{X}^{map}_m\} \cup \left\{{}^{w}\mathbf{X}^{cmc}_{a,f,o}\right\} \\ = &\{{}^{w}\mathbf{X}_m | \; \forall m \in \mathcal{M} \cup \{\mathcal{F} \times \mathcal{O}\}\}, \end{aligned} \tag{29}$$

To complete our factor graph illustrated in Fig. 1, we further define a error function $\varepsilon_{\text{a2e}}(\cdot)$. It is designed to model the dis-

crepancy between agent's top keypoints $\left\{{}^{w}\mathbf{X}^{cmc}_{a,f,o_t}\right\}$ and ego-camera positions $\{{}^{w}\mathbf{X}_{c_e,f}\}$. Since $\left\{{}^{w}\mathbf{X}^{cmc}_{a,f,o_t}\right\}$ and $\{{}^{w}\mathbf{X}_{c_e,f}\}$ are correlated, we solely apply a kinematic constraint $\varepsilon_{\text{kin}}(\cdot)$ to $\left\{{}^{w}\mathbf{X}^{cmc}_{a,f,o}\right\}$ to jointly model kinematic factors of mobile agent and ego camera. Within the kinematic factors, we hypothesize that the agent's motion vector, and, the vector from the agent's bottom (o_b) to top (o_t), may exhibit minimal changes between successive frames. The formulas are described as follows:

$$\varepsilon_{\text{a2e}}\left({}^{w}\mathbf{X}_{c_e,f}, {}^{w}\mathbf{X}^{cmc}_{a,f,o_t}\right) = \left\|{}^{w}\mathbf{X}_{c_e,f} - {}^{w}\mathbf{X}^{cmc}_{a,f,o_t}\right\|_2, \tag{30}$$

$$\begin{aligned}\varepsilon_{\text{kin}}\left({}^{w}\mathbf{X}^{cmc}_{a,f-1:f+1}\right) = &\left\|\left({}^{w}\mathbf{X}^{cmc}_{a,f+1,o_t} - {}^{w}\mathbf{X}^{cmc}_{a,f,o_t}\right)\right.\\ &\left.-\left({}^{w}\mathbf{X}^{cmc}_{a,f,o_t} - {}^{w}\mathbf{X}^{cmc}_{a,f-1,o_t}\right)\right\|_2\\ &+\left\|\left({}^{w}\mathbf{X}^{cmc}_{a,f+1,o_t} - {}^{w}\mathbf{X}^{cmc}_{a,f+1,o_b}\right)\right.\\ &\left.-\left({}^{w}\mathbf{X}^{cmc}_{a,f,o_t} - {}^{w}\mathbf{X}^{cmc}_{a,f,o_b}\right)\right\|_2,\end{aligned} \tag{31}$$

Note that, we apply a simple kinematics constraint in Eq. 31 to simplify the Jacobian matrix calculation in our custom factors [39]. However, to better formulate the kinematics of the mobile agent, we may incorporate the IMUs data [48], [112] and consider agent's joint angles and limb lengths [113], [114]. We leave these to the future works.

The factor graph [39], [40] is a graphical model that represents the probabilistic relationships between variables, observations, and constraints, facilitating efficient optimization in the scene mapping and camera pose estimation process. Given all observations $\mathcal{Z}$, the posterior probability of all variables $\mathcal{Y}$ is represented as $\mathbb{P}(\mathcal{Y}|\mathcal{Z})$. In our factor graph optimization, we want to maximize $\mathbb{P}(\mathcal{Y}|\mathcal{Z})$, which is equal to minimize $-\log(\mathbb{P}(\mathcal{Y}|\mathcal{Z}))$. This can be further simplified and expressed as

$$\begin{aligned}\arg\min_{\mathcal{Y}} \Bigg\{ &\sum_{\forall c_i \in \mathcal{C}^{cmc}\cup\mathcal{F}} \sum_{\forall f\in\mathcal{F}} \sum_{\forall m\in\mathcal{M}\cup\{\mathcal{F}\times\mathcal{O}\}} \|\varepsilon_{\text{c2k}}\|^2_{\Sigma_{\text{c2k}}}\\ &+ \sum_{\forall f\in\mathcal{F}} \|\varepsilon_{\text{kin}}\|^2_{\Sigma_{\text{kin}}} + \sum_{\forall f\in\mathcal{F}} \|\varepsilon_{\text{a2e}}\|^2_{\Sigma_{\text{a2e}}} \Bigg\},\\ \text{where } \mathcal{Y} = &\{\mathcal{T}, \mathcal{X}\}\\ \mathcal{Z} = &\{\mathbf{x}_{c_i,f,m} \mid \forall c_i \in \mathcal{C}^{cmc}\cup\mathcal{F},\ \forall f\in\mathcal{F},\\ &\forall m\in\mathcal{M}\cup\{\mathcal{F}\times\mathcal{O}\}\}.\end{aligned} \tag{32}$$

where Σ_{c2k}, Σ_{a2e} and Σ_{kin} denote the covariance matrices.

Using optimized ego-camera poses and corresponding RGB-D images, we refine the constructed point clouds. We then apply the planar patch detection [88] on refined point clouds to obtain a parameterized scene layout (Fig. 3). Our target outputs, as the parameterized scene layout and CMC 6-DoF poses registered to it, have been achieved.

D. Register 6-DoF Poses for Isolated CMCs

Although previous processes require at least two overpping-view CMCs to estimate CMC relative 6-DoF poses and pseudo-scale agent keypoint trajectories, there may be isolated CMC that has no overlapping views with others. Despite this,

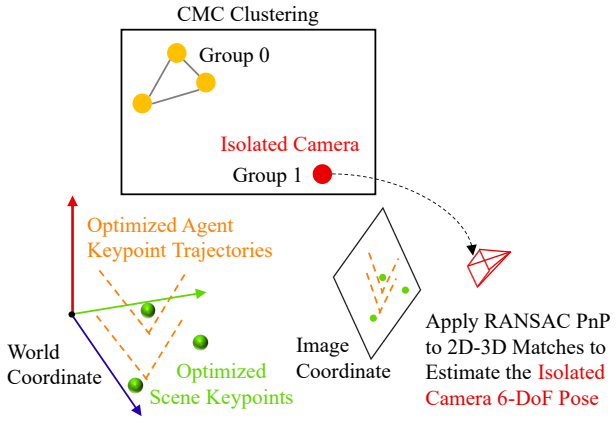

Fig. 5: **Register an isolated CMC 6-DoF pose.** This process involves the utilization of static scene keypoints and mobile agent keypoints to establish a 2D-3D matches for isolated CMCs. Following this, PnP and RANSAC are performed to estimate the isolated camera 6-DoF pose.

YOWO can still register such an isolated CMC to the scene layout. Fig. 5 gives an example, of how YOWO registers an isolated CMC to the scene layout. Basically, YOWO processes the isolated CMC after completing the scene mapping and non-isolated CMC registration. It is assumed that the isolated CMC can observe agent mobile keypoints and scene static keypoints, which have been jointly optimized by YOWO in the preceding steps. By referring to the timestamps of constructed 3D mobile keypoints, we can form the 2D-3D matches for observed mobile keypoints. Employing RANSAC and PnP [99], [106], [107], we initialize the 6-DoF pose of the isolated CMC. Based the initialized 6-DoF pose prior and visual features, we further obtain 2D-3D matches for observed scene static keypoints. We then incorporate the 2D-3D matches of both scene static keypoints and mobile keypoints to refine the isolated CMC 6-DoF pose.

IV. Experiments

A. Experimental Setup

Dataset. As shown in Tab.I, we contribute a CMC and ego-camera collaborative capturing dataset for jointly evaluating scene layout mapping and CMC 6-DoF pose registration. Utilizing AI2THOR [41], [42] and Gym-UnrealCV [43], [44], we generate indoor scenes that encompass the house, garage, school, etc. An agent is then navigated within these scenes for collaborative capture. Taking advantage of simulation, we introduce variations in texture, illumination, object presence, and scale within these indoor scenes. This ensures a comprehensive evaluation of our proposed method under a wide array of conditions.

Metrics. We utilize standard metrics commonly used in visual localization [15], [25], [27], [60] to evaluate CMC 6-DoF poses. The valuation metrics are represented by

$$\varepsilon_{\text{pos}} = \|\mathbf{X}_{c_i} - \mathbf{X}^{\text{gt}}_{c_i}\|_2 \tag{33}$$

$$\varepsilon_{\text{rot}} = \arccos\left(\frac{\text{trace}\left(\mathbf{R}^{\top}_{c_i}\cdot\mathbf{R}^{\text{gt}}_{c_i}\right)-1}{2}\right) \tag{34}$$

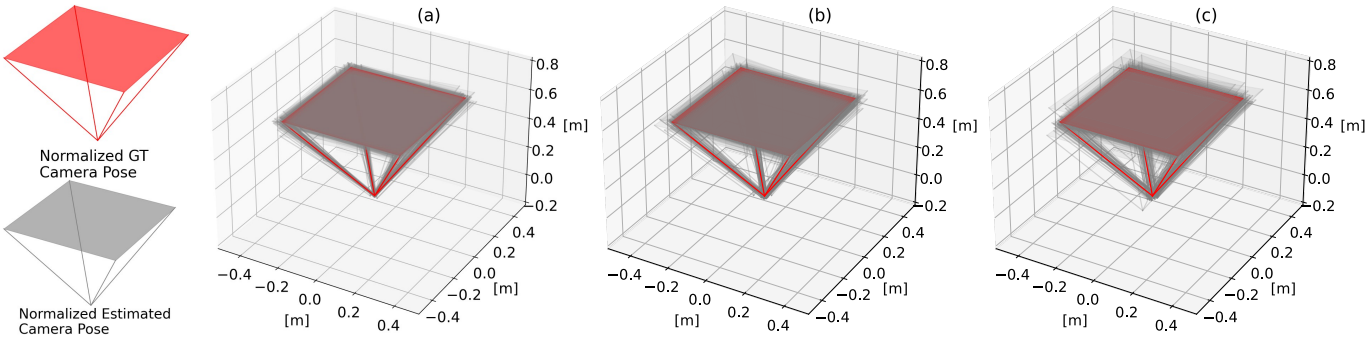

Fig. 6: **Visualization of bias between estimated CMC 6-DoF poses and GTs, in a normalized space with all GTs aligned to one pose.** The normalized GT and estimated camera poses are rendered with red and gray colors, respectively. The relative 6-DoF pose estimation bias of methods YOWO, OMECC [13], and ECCMP [12] are plotted in (a), (b), and (c), respectively.

Besides, to evaluate the scene mapping performance, we follow [50], [75], [115] to compare the estimated ego-camera trajectory to GT using ATE RMSE, and follow [116]–[118] to evaluate the 2D top-view scene layout with polygon IoU.

B. Comparison with State-of-the-Art (SOTA) Methods

YOWO vs. SOTA CMC relative 6-DoF pose estimation. ECCMP [12] and OMECC [13] are representative works that utilize mobile keypoints for CMC relative 6-DoF pose estimation. Although their estimated CMC relative 6-DoF poses are not correlated with the scene layout and are isolated within each room, we follow their evaluation protocol to align CMC relative 6-DoF poses to the GT CMC layout for the evaluation. As shown in Tab.II and Fig. 6, all methods yield promising results by leveraging the mobile keypoints. Our YOWO method, however, outperforms all others in every metric, albeit the improvement is marginal. The rationale here is that mobile keypoints provide robust cues for cross-camera matching, which plays a vital role in CMC relative 6-DoF pose estimation. Besides, due to the visual ambiguity (Fig. 2), even though YOWO further incorporates static scene keypoints into its joint optimization, the improvement is marginal compared to ECCMP [12] and OMECC [13] that solely rely on mobile keypoints. Nonetheless, it is important to note that, the CMC relative 6-DoF poses possess a pseudo scale and are not automatically aligned with the scene layout, hindering their practical application. Hence, it is recommended to move to the CMC 6-DoF pose registration, which requires automatically aligning CMC 6-DoF poses to the scene layout.

YOWO vs. SOTA scene mapping. We compare YOWO with a traditional RGB-D SLAM method—BAD-SLAM [47], and a more recent solution—NICE-SLAM [50], which utilizes neural implicit representation. For all methods, we apply a planar detector [88] to parameterize obtained point clouds to scene layouts. The results are presented in Tab. V. Unlike other visual localization datasets [51], [75], our dataset features a fast-moving agent that may not form loop closures, which often leads to drift errors when RGB-D SLAM methods are employed. While BAD-SLAM performs on par with NICE-SLAM, YOWO outperforms both, benefiting from the incorporation of CMC observations to enhance the ego-camera scene mapping. Although we evaluate scene mapping performance,

TABLE II: **Comparison with SOTAs on CMC relative 6-DoF pose estimation.** The relative 6-DoF poses are aligned to the GT CMC layout by using the KU algorithm [109], [110]. The identical mobile keypoints are used for all methods. The best and the second-best results in each metric are highlighted in **bold** and underlined, respectively.

Method	ε_{pos} [m] ↓			ε_{rot} [deg] ↓		
	Avg.	Min.	Max.	Avg.	Min.	Max.
ECCMP	0.024	**0.002**	0.105	0.338	0.097	0.811
OMECC	0.015	**0.002**	0.062	0.216	0.103	0.529
YOWO	**0.008**	**0.002**	**0.033**	**0.135**	**0.092**	**0.359**

TABLE V: **Comparison with SOTAs on scene mapping.** The best and the second-best results in each metric are highlighted in **bold** and underlined, respectively.

	Ego-camera ATE RMSE [m] ↓	Layout IoU ↑
BAD-SLAM	0.286	0.848
NICE-SLAM	0.252	0.873
YOWO	**0.122**	**0.927**

YOWO is not tied to a specific SLAM method and can adapt to incorporate more advanced SLAM methods in the future.

YOWO vs. SOTA CMC 6-DoF pose registration. We compare YOWO with SOTA visual localization methods: Hloc [15] and Kapture [25], [56]. For a fair comparison, all methods utilize our estimated ego-camera poses to construct the identical real-scale 3D scenes, as opposed to employing pseudo-scale ego-camera poses estimated by COLMAP [46]. Notably, the scene reconstruction bias is inherently transferred to the CMC 6-DoF pose registration, aligning more closely with real-world application scenarios.

As shown in Tab. III, using the identical 3D scene model, YOWO delivers a performance leap over Hloc and Kapture on CMC 6-DoF pose registration. Both Hloc and Kapture have their maximum ε_{pos} and ε_{rot} exceeding 10 meters and 50 degrees, respectively. In contrast, YOWO maintains these errors within 0.4 meters and 0.8 degrees, respectively. For Hloc and Kapture, the quality of CMC 6-DoF pose registration heavily relies on global search (*i.e.*, image retrieval). However,

TABLE III: **Comparison with SOTAs on CMC 6-DoF pose registration.** Since the estimated scene layouts have been aligned to their GTs, the registered CMC 6-DoF poses are directly compared to their GTs. The best results in each metric are highlighted in **bold**.

Method	ε_{pos} [m] ↓			ε_{rot} [deg] ↓		
	Avg.	Min.	Max.	Avg.	Min.	Max.
Hloc	1.345	0.014	11.132	14.958	0.367	54.231
Kapture	1.187	0.011	10.787	12.635	0.236	57.958
YOWO	**0.131**	**0.007**	**0.348**	**0.205**	**0.026**	**0.762**

TABLE IV: **Evaluation of SOTAs using YOWO's CMC 6-DoF pose priors to assist their 2D-3D matches.** By using YOWO CMC 6-DoF pose priors and the 3D scene model, Hloc and Kapture can circumvent the vulnerable global search and effectively construct 2D-3D matches.

Method	ε_{pos} [m] ↓			ε_{rot} [deg] ↓		
	Avg.	Min.	Max.	Avg.	Min.	Max.
Hloc	0.259	0.010	0.678	0.711	0.149	1.247
Kapture	0.213	0.008	0.634	0.585	0.122	1.321

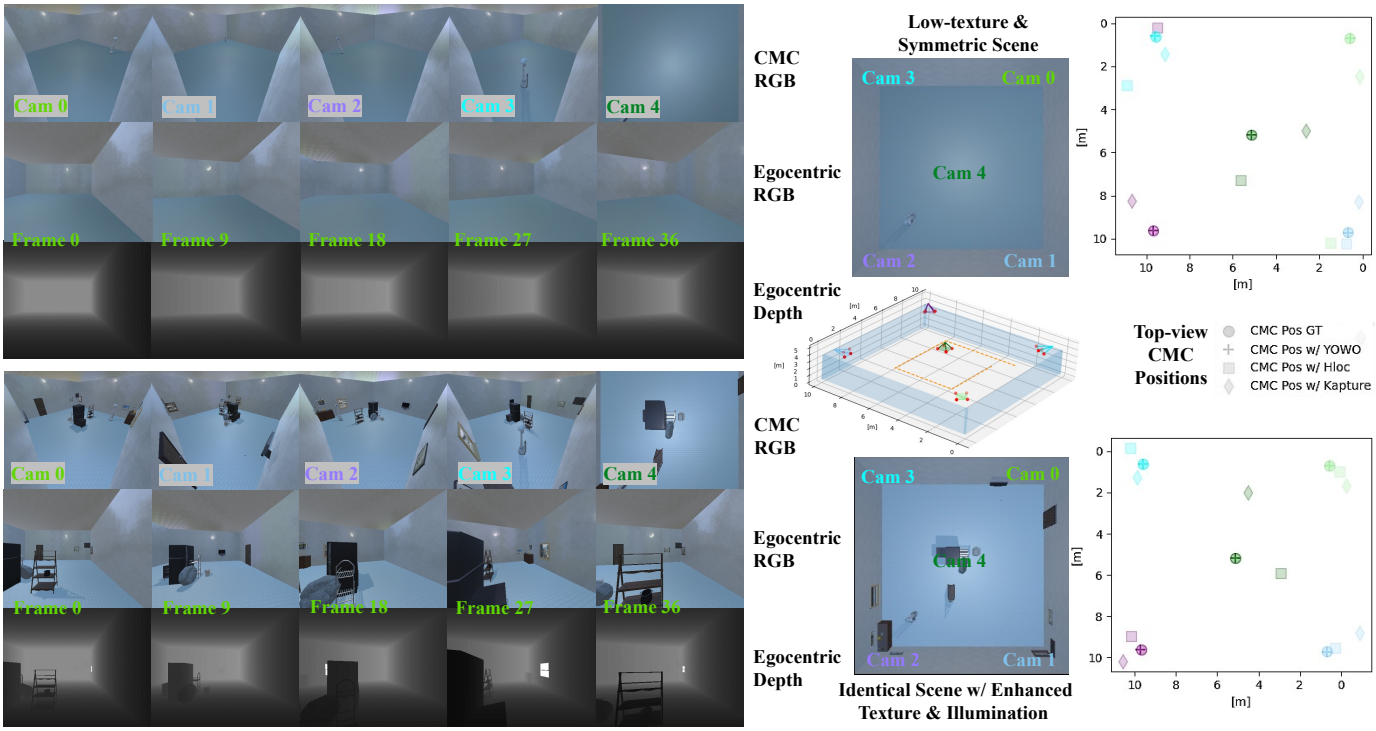

Fig. 7: **Comparison between YOWO, Hloc [15], and Kapture [25], [56] for CMC 6-DoF pose registration.** Using the identical room, we specifically focus on analyzing the effect of various illumination and texture, and further infer the robustness of each model.

TABLE VI: **Ablation studies on YOWO.** See Fig. 3, three processes of YOWO include: (i) ego-camera processing, (ii) CMC processing, and (iii) collaborative processing. JO denotes the joint optimization with a factor graph using Eq. 32.

	CMC registration		Scene mapping	
	Avg. ε_{pos} [m] ↓	Avg. ε_{rot} [deg] ↓	Ego-camera ATE RMSE [m] ↓	Layout IoU ↑
(i)	-	-	0.165	0.915
(i)(ii)(iii) w/o JO	0.173	0.318	0.165	0.915
(i)(ii)(iii) w/ JO	0.131 (-0.042)	0.205 (-0.113)	0.122 (-0.043)	0.927 (+0.012)

to make their global search work properly, the query CMC viewpoint should be similar to one of the ego-camera reference viewpoints, which is challenging as the CMC positions may not be accessible to the ego camera (Fig. 2). Therefore, remarkable errors are reported in Tab. III. Moreover, those errors could potentially increase when applied to larger indoor scenes.

We also conduct additional experiments, as detailed in Tab. IV, to investigate the feasibility of applying YOWO's CMC 6-DoF pose priors to assist 2D-3D matches for Hloc and Kapture. The results indicate that the performance of Hloc and Kapture improves considerably, nearly matching the level of YOWO. This reveals that YOWO's CMC 6-DoF pose priors can bestow enhanced robustness upon existing visual localization methods. However, unlike YOWO, Hloc and Kapture are unable to include mobile keypoints to perform a joint

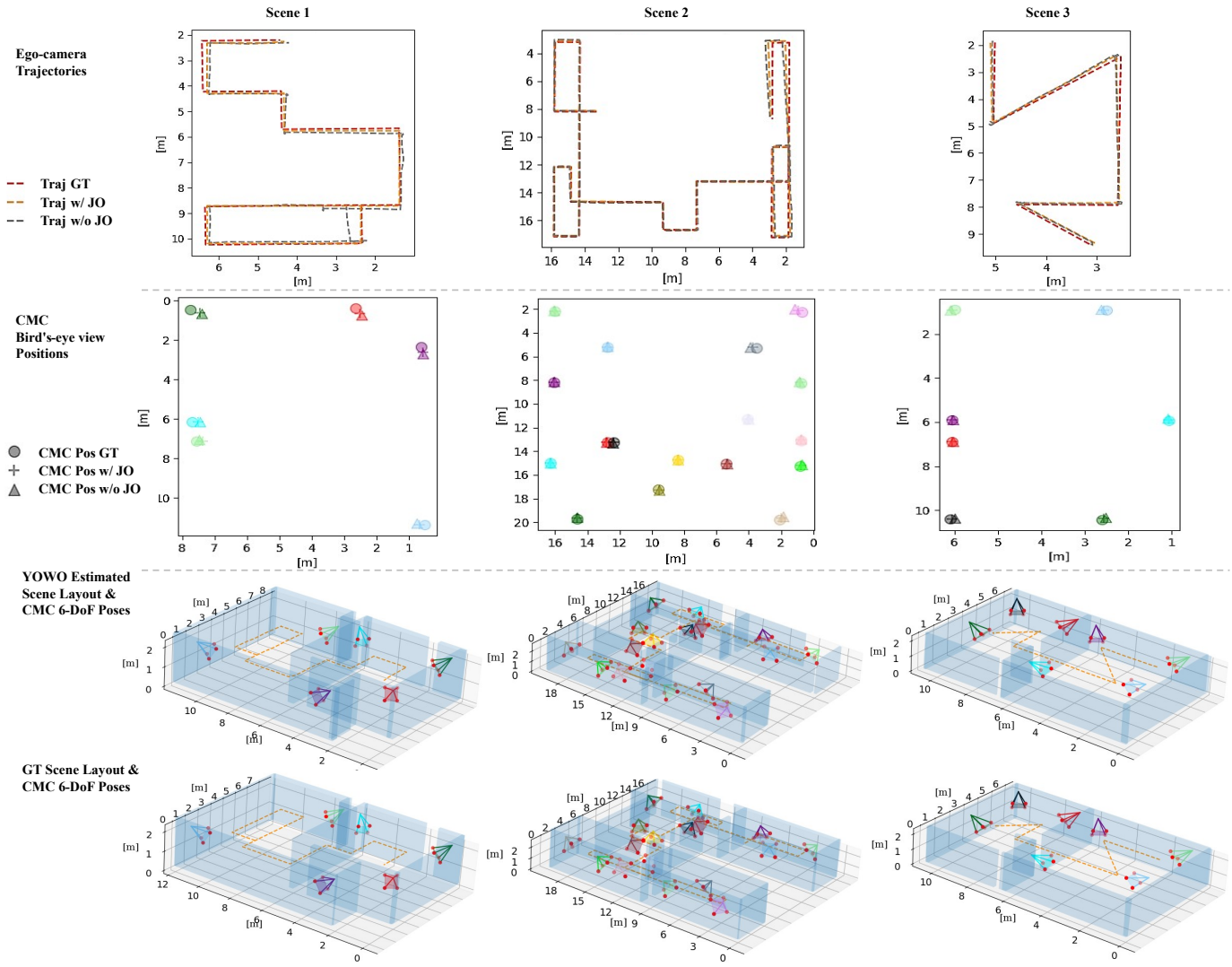

Fig. 8: **Visualization of ablation study results.** Three indoor scenes are illustrated as examples. In each scene, we plot ego-camera trajectories and CMC positions for the GT, the estimations w/ JO, and the estimations w/o JO, respectively.

optimization with scene static keypoints. As a result, even with YOWO's CMC 6-DoF pose priors, Hloc and Kapture still exhibit slightly weaker performance than YOWO.

In the end, we include examples in Fig. 7 to explore how the scene appearance affect the robustness of each model. In an indoor scene with a highly symmetrical structure and ambiguous texture, our YOWO can obtain accurate 6-DoF CMC poses by leveraging mobile keypoints. On the other hand, such a challenging scene can easily cause conventional visual localization to fail, as it is difficult to correctly determine the correspondence between ego-camera and CMC captures by solely relying on the static scene features. After introducing new objects and adjusting the illumination, the static scene features became more distinctive. Consequently, the performance of traditional visual localization methods has seen significant improvement. However, for conventional visual localization to work properly, the viewpoint of the observed CMC must be similar to one of the ego-camera's reference viewpoints, which is difficult because the CMC's positions may not be accessible to the ego-camera. As a result, we can still observe notable biases in some of their results. In a nutshell, YOWO can generate high-performance results in both scenes, which reveals the robustness of YOWO.

C. Ablation Studies

We conduct ablation studies to validate the effectiveness of joint optimization in YOWO. The scene mapping performance is initially assessed using only ego-camera processing. Subsequently, the CMC 6-DoF registration is evaluated by directly aligning CMC processing results with the raw scene layout. Finally, the contribution of jointly optimizing scene mapping and CMC 6-DoF registration using a factor graph is evaluated.

While we have shown that unifying scene mapping and CMC 6-DoF pose registration in a single framework overcomes existing challenges (Tabs. II,V,III,IV), the results presented in Tab. VI and Fig. 8 further demonstrate that customizing a factor graph for joint optimization can narrow the gap between those two tasks and enhance the overall performance. Through jointly optimizing scene mapping and CMC 6-DoF pose registration, both agent trajectories and CMC positions

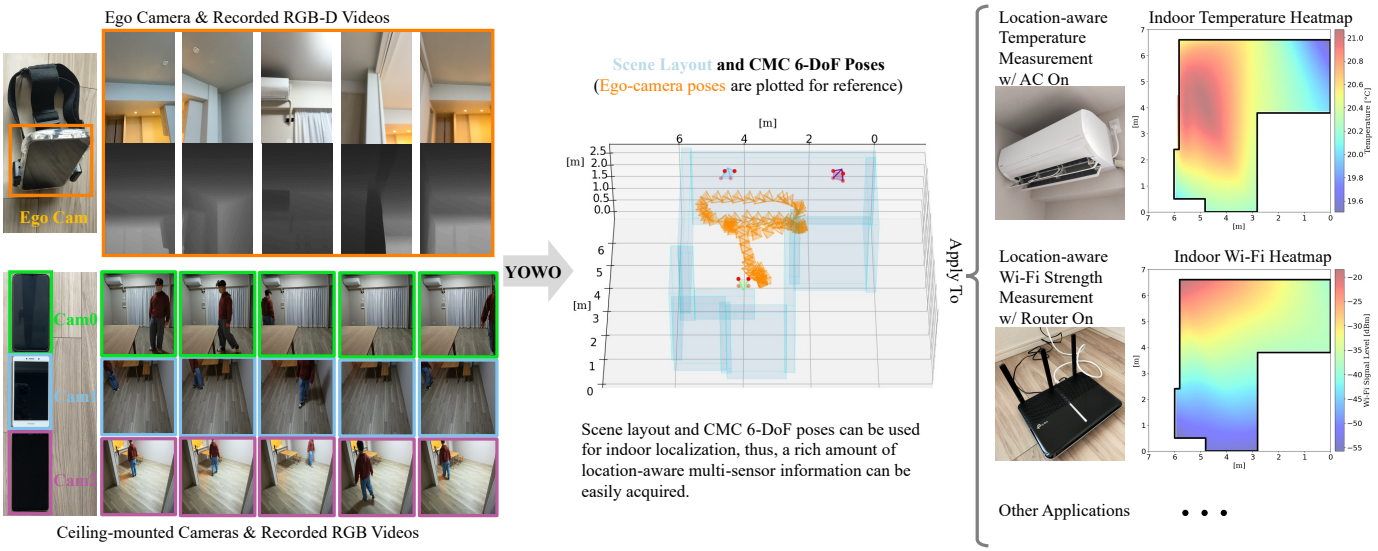

Fig. 9: **Downstream applications of YOWO in real-life environments.** We capture data in a real indoor scene with four synchronized phone cameras: one for the ego camera and three for CMCs. We then apply YOWO to obtain the scene layout and register CMCs to the scene, which facilitate the collection of position-aware multi-sensor data for downstream applications.

more closely align with the ground truths in listed examples of Fig. 8. Consequently, the estimated scene layouts are also presented appropriate structures as illustrated. To this end, the overall effectiveness and efficiency of our YOWO has been verified.

D. Proof-of-concept Downstream Experiments

The results produced by YOWO, including scene layout and registered CMCs, have the potential to advance a plethora of downstream studies. The direct application of YOWO delivers the capability to automatically model scenes and CMCs, thereby enhancing the efficiency of previous research that employ CMCs for indoor visual observation. It can also assist with indoor scene mapping tasks. For intermediate applications, calibrated CMCs can be leveraged to estimate the 3D object pose under the multi-camera triangulation scheme [85]. This allows for efficient collection of position-aware multi-sensor data in a challenging indoor scene. Consequently, we are capable of supplying 3D body pose labels, encompassing pose estimation, trajectory estimation, and action recognition [9], [119], [120].

Additionly, we provide proof-of-concept experiments in Fig. 9. These experiments extend the downstream applications of YOWO beyond the realm of computer vision domain, demonstrating its applicability to other forms of sensing such as Wi-Fi signals and temperature measurements. Note that, some potential downstream applications of YOWO might involve using CMCs to capture images of individuals in public spaces. We advocate for careful ethical considerations in such scenarios.

V. Discussion

Limitations and applicable scopes. (i) YOWO works on the assumption of a single agent in the scene, which limits its applicability in indoor scenes with multiple agent-like entities. Thus, when the mobile agent is a person, YOWO is best utilized in people-free settings, such as a supermarket after business hours. While this may seem limiting, it offers two key benefits. Firstly, it avoids the potentially offensive use of an ego camera to capture unknown individuals; secondly, it mitigates the risk of cross-camera association errors. (ii) In YOWO, the performance of the ego-camera SLAM and the CMC relative pose estimation are affected by the agent's path. It determines the scene captures for the ego-camera SLAM, and the spatiotemporal distribution of mobile keypoints for the CMC processing. By examining YOWO results in Tabs. II, V, and III, we deduce that the ego-camera SLAM primarily influences the CMC registration bias. Therefore, additional efforts are required to plan the agent's path. The objective is to increase the overlapping score as outlined in Eqs. 7 and 8 for the CMC processing, and to promote loop closures for the ego-camera SLAM. Since the outputs of YOWO primarily serve downstream applications and YOWO accommodates offline processing, it is feasible to adopt an iterative approach for optimizing the path planning. Specifically, we can execute YOWO multiple times, employing a feedback loop where initial results from the 3D scene modeling and CMC observations are used to guide subsequent adjustments to the capture path.

Conclusion and potential impacts. We introduce YOWO, the first-of-its-kind framework that can jointly map an indoor scene and register CMCs to the scene layout. The CMCs could be either Closed-Circuit Television (CCTV) cameras or simply smartphone cameras, providing flexibility for use in various indoor environments. Additionally, we propose an inaugural dataset to reveal the synergy between ceiling-mounted cameras and the ego camera, establishing the first benchmark for collaborative scene mapping and CMC 6-DoF registration. Leveraging mobile keypoints of a mobile agent, YOWO surmounts the visual ambiguity challenge in visual localization, and reduces the uncertainty inherent in indoor

SLAM. Hence, YOWO underpins an efficient and robust solution for indoor mapping and 3D visual positioning.

REFERENCES

[1] Z. Zhou, X. Chen, Y.-C. Chung, Z. He, T. X. Han, and J. M. Keller, "Activity analysis, summarization, and visualization for indoor human activity monitoring," IEEE transactions on circuits and systems for video technology, vol. 18, no. 11, pp. 1489–1498, 2008.

[2] H. Aghajan and A. Cavallaro, Multi-camera networks: principles and applications. Academic press, 2009.

[3] K. Li, Q. Dai, and W. Xu, "Markerless shape and motion capture from multiview video sequences," IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 3, pp. 320–334, 2011.

[4] X. Wang, "Intelligent multi-camera video surveillance: A review," Pattern recognition letters, vol. 34, no. 1, pp. 3–19, 2013.

[5] J. Xu, S. Denman, S. Sridharan, and C. Fookes, "An efficient and robust system for multiperson event detection in real-world indoor surveillance scenes," IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 6, pp. 1063–1076, 2014.

[6] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "A compressive sensing approach to describe indoor scenes for blind people," IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 7, pp. 1246–1257, 2014.

[7] Q. Shi, S. Nobuhara, and T. Matsuyama, "Augmented motion history volume for spatiotemporal editing of 3-d video in multiparty interaction scenes," IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 1, pp. 63–76, 2015.

[8] W. Li, Y. Wong, A.-A. Liu, Y. Li, Y.-T. Su, and M. Kankanhalli, "Multi-camera action dataset for cross-camera action recognition benchmarking," in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, pp. 187–196.

[9] Y. Dai, C. Wen, H. Wu, Y. Guo, L. Chen, and C. Wang, "Indoor 3d human trajectory reconstruction using surveillance camera videos and point clouds," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 4, pp. 2482–2495, 2021.

[10] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21 096–21 106.

[11] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, Y. Yao, L. Zheng, M. S. Rahman, M. S. Arya, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, S. Prajapati, A. Li, S. Li, K. Kunadharaju, S. Jiang, and R. Chellappa, "The 7th ai city challenge," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2023.

[12] S.-E. Lee, K. Shibata, S. Nonaka, S. Nobuhara, and K. Nishino, "Extrinsic camera calibration from a moving person," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 10 344–10 351, 2022.

[13] B. Pätzold, S. Bultmann, and S. Behnke, "Online marker-free extrinsic camera calibration using person keypoint detections," in DAGM German Conference on Pattern Recognition. Springer, 2022, pp. 300–316.

[14] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7199–7209.

[15] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 716–12 725.

[16] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, and S. Tang, "Egobody: Human body shape and motion of interacting people from head-mounted devices," in European conference on computer vision (ECCV), Oct. 2022.

[17] R. Khirodkar, A. Bansal, L. Ma, R. Newcombe, M. Vo, and K. Kitani, "Egohumans: An egocentric 3d multi-human benchmark," arXiv preprint arXiv:2305.16487, 2023.

[18] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 11, pp. 1330–1334, 2000.

[19] A. Schmidt, A. Kasiński, M. Kraft, M. Fularz, and Z. Domagała, "Calibration of the multi-camera registration system for visual navigation benchmarking," International Journal of Advanced Robotic Systems, vol. 11, no. 6, p. 83, 2014.

[20] M. Munaro, F. Basso, and E. Menegatti, "Openptrack: Open source multi-camera calibration and people tracking for rgb-d camera networks," Robotics and Autonomous Systems, vol. 75, pp. 525–538, 2016.

[21] F. Rameau, J. Park, O. Bailo, and I. S. Kweon, "Mc-calib: A generic and robust calibration toolbox for multi-camera systems," Computer Vision and Image Understanding, vol. 217, p. 103353, 2022.

[22] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 1052–1067, 2007.

[23] E. Ataer-Cansizoglu, Y. Taguchi, S. Ramalingam, and Y. Miki, "Calibration of non-overlapping cameras using an external slam system," in 2014 2nd International Conference on 3D Vision, vol. 1. IEEE, 2014, pp. 509–516.

[24] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2938–2946.

[25] M. Humenberger, Y. Cabon, N. Pion, P. Weinzaepfel, D. Lee, N. Guérin, T. Sattler, and G. Csurka, "Investigating the role of image retrieval for visual localization: An exhaustive benchmark," International Journal of Computer Vision, vol. 130, no. 7, pp. 1811–1836, 2022.

[26] P.-E. Sarlin, M. Dusmanu, J. L. Schönberger, P. Speciale, L. Gruber, V. Larsson, O. Miksik, and M. Pollefeys, "Lamar: Benchmarking localization and mapping for augmented reality," in European Conference on Computer Vision. Springer, 2022, pp. 686–704.

[27] T. Do, O. Miksik, J. DeGol, H. S. Park, and S. N. Sinha, "Learning to detect scene landmarks for camera localization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11 132–11 142.

[28] V. Panek, Z. Kukelova, and T. Sattler, "Meshloc: Mesh-based visual localization," in European Conference on Computer Vision. Springer, 2022, pp. 589–609.

[29] Z. Wang, L. Zhang, S. Zhao, and Y. Zhou, "Ct-lvi: A framework towards continuous-time laser-visual-inertial odometry and mapping," IEEE Transactions on Circuits and Systems for Video Technology, pp. 1–1, 2023.

[30] H. Liu, L. Zhao, Z. Peng, W. Xie, M. Jiang, H. Zha, H. Bao, and G. Zhang, "A low-cost and scalable framework to build large-scale localization benchmark for augmented reality," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2274–2288, 2024.

[31] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys, "Joint camera pose estimation and 3d human pose estimation in a multi-camera setup," in Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12. Springer, 2015, pp. 473–487.

[32] Y. Xu, Y.-J. Li, X. Weng, and K. Kitani, "Wide-baseline multi-camera calibration using person re-identification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 134–13 143.

[33] B. Huang, Y. Shu, T. Zhang, and Y. Wang, "Dynamic multi-person mesh recovery from uncalibrated multi-view cameras," in 2021 International Conference on 3D Vision (3DV). IEEE, 2021, pp. 710–720.

[34] J. L. Charco, A. D. Sappa, B. X. Vintimilla, and H. O. Velesaca, "Camera pose estimation in multi-view environments: From virtual scenarios to the real world," Image and Vision Computing, vol. 110, p. 104182, 2021.

[35] K. Liu, L. Chen, L. Xie, J. Yin, S. Gan, Y. Yan, and E. Yin, "Auto calibration of multi-camera system for human pose estimation," IET Computer Vision, vol. 16, no. 7, pp. 607–618, 2022.

[36] B. Gordon, S. Raab, G. Azov, R. Giryes, and D. Cohen-Or, "Flex: Extrinsic parameters-free multi-view 3d human motion reconstruction," in European Conference on Computer Vision. Springer, 2022, pp. 176–196.

[37] N. Saini, C.-H. P. Huang, M. J. Black, and A. Ahmad, "Smartmocap: Joint estimation of human and camera motion using uncalibrated rgb cameras," IEEE Robotics and Automation Letters, 2023.

[38] H. Ci, M. Liu, X. Pan, fangwei zhong, and Y. Wang, "Proactive multi-camera collaboration for 3d human pose estimation," in The Eleventh International Conference on Learning Representations, 2023. [Online]. Available: https://openreview.net/forum?id=CPIy9TWFYBG

[39] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," Georgia Institute of Technology, Tech. Rep, vol. 2, p. 4, 2012.

[40] F. Dellaert, M. Kaess et al., "Factor graphs for robot perception," Foundations and Trends® in Robotics, vol. 6, no. 1-2, pp. 1–139, 2017.

[41] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu et al., "Ai2-thor: An interactive 3d environment for visual ai," arXiv preprint arXiv:1712.05474, 2017.

[42] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," in NeurIPS, 2022.

[43] F. Zhong, W. Qiu, T. Yan, A. Yuille, and Y. Wang, "Gym-unrealcv: Realistic virtual worlds for visual reinforcement learning," 2017. [Online]. Available: https://github.com/unrealcv/gym-unrealcv

[44] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, and Y. Wang, "Unrealcv: Virtual worlds for computer vision," in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1221–1224.

[45] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," International journal of computer vision, vol. 80, pp. 189–210, 2008.

[46] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4104–4113.

[47] T. Schops, T. Sattler, and M. Pollefeys, "Bad slam: Bundle adjusted direct rgb-d slam," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[48] 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.

[49] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "Neuralrecon: Real-time coherent 3d reconstruction from monocular video," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15 598–15 607.

[50] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 786–12 796.

[51] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5828–5839.

[52] Y. Mao, Y. Zhang, H. Jiang, A. Chang, and M. Savva, "Multiscan: Scalable rgbd scanning for 3d environments with articulated objects," Advances in Neural Information Processing Systems, vol. 35, pp. 9058–9071, 2022.

[53] Y. Zheng, Y. Yang, K. Mo, J. Li, T. Yu, Y. Liu, C. K. Liu, and L. J. Guibas, "Gimo: Gaze-informed human motion prediction in context," in European Conference on Computer Vision. Springer, 2022, pp. 676–694.

[54] X. Yi, Y. Zhou, M. Habermann, V. Golyanik, S. Pan, C. Theobalt, and F. Xu, "Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors," ACM Transactions on Graphics (TOG), vol. 42, no. 4, 2023.

[55] M. Ding, Z. Wang, J. Sun, J. Shi, and P. Luo, "Camnet: Coarse-to-fine retrieval for camera re-localization," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2871–2880.

[56] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, V. Leroy, J. Revaud, P. Rerole, N. Pion, C. de Souza, and G. Csurka, "Robust image retrieval-based visual localization using kapture," arXiv preprint arXiv:2007.13867, 2020.

[57] E. Brachmann and C. Rother, "Visual camera re-localization from rgb and rgb-d images using dsac," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 9, pp. 5847–5865, 2021.

[58] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 224–236.

[59] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8922–8931.

[60] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic et al., "Benchmarking 6dof outdoor visual localization in changing conditions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8601–8610.

[61] E. Spera, A. Furnari, S. Battiato, and G. M. Farinella, "Egocart: A benchmark dataset for large-scale indoor image-based localization in retail stores," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 4, pp. 1253–1267, 2019.

[62] J. Wald, T. Sattler, S. Golodetz, T. Cavallari, and F. Tombari, "Beyond controlled environments: 3d camera re-localization in changing indoor scenes," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. Springer, 2020, pp. 467–487.

[63] E. Arnold, J. Wynn, S. Vicente, G. Garcia-Hernando, A. Monszpart, V. Prisacariu, D. Turmukhambetov, and E. Brachmann, "Map-free visual relocalization: Metric pose relative to a single image," in European Conference on Computer Vision. Springer, 2022, pp. 690–708.

[64] S. Chen, X. Li, Z. Wang, and V. A. Prisacariu, "Dfnet: Enhance absolute pose regression with direct feature matching," in European Conference on Computer Vision. Springer, 2022, pp. 1–17.

[65] Y. Matsumoto, G. Nakano, and K. Ogura, "Indoor visual localization using point and line correspondences in dense colored point cloud," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 3616–3625.

[66] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena, "Leveraging deep visual descriptors for hierarchical efficient localization," in Conference on Robot Learning. PMLR, 2018, pp. 456–465.

[67] E. Brachmann and C. Rother, "Expert sample consensus applied to camera re-localization," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7525–7534.

[68] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14 141–14 152.

[69] L. Su, Y. Lin, and L. Zhang, "Multi-camera joint self-calibration from observations of pedestrians," in Proceedings of the 2021 4th International Conference on Image and Graphics Processing, 2021, pp. 120–124.

[70] W.-C. Ma, A. J. Yang, S. Wang, R. Urtasun, and A. Torralba, "Virtual correspondence: Humans as a cue for extreme-view geometry," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15 924–15 934.

[71] S. Ardeshir and A. Borji, "Ego2top: Matching viewers in egocentric and top-view videos," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, 2016, pp. 253–268.

[72] ——, "Egocentric meets top-view," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 6, pp. 1353–1366, 2018.

[73] J. P. Araújo, J. Li, K. Vetrivel, R. Agarwal, J. Wu, D. Gopinath, A. W. Clegg, and K. Liu, "Circle: Capture in rich contextual environments," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21 211–21 221.

[74] S. Bultmann, R. Memmesheimer, and S. Behnke, "External camera-based mobile robot pose estimation for collaborative perception with smart edge sensors," 2023.

[75] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE, 2012, pp. 573–580.

[76] J. Li, J. Xu, F. Zhong, X. Kong, Y. Qiao, and Y. Wang, "Pose-assisted multi-camera collaboration for active object tracking," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 759–766.

[77] G. Bradski, "The opencv library." Dr. Dobb's Journal: Software Tools for the Professional Programmer, vol. 25, no. 11, pp. 120–123, 2000.

[78] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," Image and vision computing, vol. 10, no. 3, pp. 145–155, 1992.

[79] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in Proceedings third international conference on 3-D digital imaging and modeling. IEEE, 2001, pp. 145–152.

[80] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5556–5565.

[81] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4938–4947.

[82] R. Pautrat, J.-T. Lin, V. Larsson, M. R. Oswald, and M. Pollefeys, "Sold2: Self-supervised occlusion-aware line description and detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11 368–11 378.

[83] G. Xu and Z. Zhang, Epipolar geometry in stereo, motion and object recognition: a unified approach. Springer Science & Business Media, 1996, vol. 6.

[84] L. Sun, W. Yin, E. Xie, Z. Li, C. Sun, and C. Shen, "Improving monocular visual odometry using learned depth," IEEE Transactions on Robotics, vol. 38, no. 5, pp. 3173–3186, 2022.

[85] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge university press, 2003.

[86] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings. Springer, 2000, pp. 298–372.

[87] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977. Springer, 2006, pp. 105–116.

[88] A. M. C. Araujo and M. M. Oliveira, "A robust statistics approach for plane detection in unorganized point clouds," Pattern Recognition, vol. 100, pp. 1–12, 2020, article 107115.

[89] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in European Conference on Computer Vision (ECCV), 2018.

[90] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in CVPR, 2019.

[91] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in European Conference on Computer Vision. Springer, 2022, pp. 1–21.

[92] Ultralytics, "Yolov8: Scalable object detection and tracking," 2023. [Online]. Available: https://github.com/ultralytics/yolov8

[93] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." Analytical chemistry, vol. 36, no. 8, pp. 1627–1639, 1964.

[94] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[95] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," International Journal of Computer Vision, vol. 124, no. 2, pp. 237–254, Sep. 2017.

[96] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration." VISAPP (1), vol. 2, no. 331-340, p. 2, 2009.

[97] H. Strasdat, A. J. Davison, J. M. Montiel, and K. Konolige, "Double window optimisation for constant time visual slam," in 2011 international conference on computer vision. IEEE, 2011, pp. 2352–2359.

[98] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "Magsac++, a fast, reliable and accurate robust estimator," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1304–1312.

[99] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981.

[100] A. Kushal and S. Agarwal, "Visibility based preconditioning for bundle adjustment," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 1442–1449.

[101] S. E. Schaeffer, "Graph clustering," Computer science review, vol. 1, no. 1, pp. 27–64, 2007.

[102] R. I. Hartley, "In defense of the eight-point algorithm," IEEE Transactions on pattern analysis and machine intelligence, vol. 19, no. 6, pp. 580–593, 1997.

[103] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 2599–2606.

[104] A. Rau, G. Garcia-Hernando, D. Stoyanov, G. J. Brostow, and D. Turmukhambetov, "Predicting visual overlap of images through interpretable non-metric box embeddings," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, 2020, pp. 629–646.

[105] J. Park, Q.-Y. Zhou, and V. Koltun, "Colored point cloud registration revisited," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 143–152.

[106] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in CVPR 2011. IEEE, 2011, pp. 2969–2976.

[107] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, and M. Okutomi, "Revisiting the pnp problem: A fast, general and optimal solution," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2344–2351.

[108] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in 2020 IEEE international conference on robotics and automation (ICRA). IEEE, 2020, pp. 4203–4210.

[109] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, vol. 32, no. 5, pp. 922–923, 1976.

[110] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 13, no. 04, pp. 376–380, 1991.

[111] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 86–97, 2012.

[112] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," IEEE Transactions on Robotics, vol. 28, pp. 61–76, 2012.

[113] T. McGrath and L. Stirling, "Body-worn imu human skeletal pose estimation using a factor graph-based optimization framework," Sensors, vol. 20, no. 23, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/6887

[114] S. Bultmann and S. Behnke, "Real-time multi-view 3d human pose estimation using semantic feedback to smart edge sensors," arXiv preprint arXiv:2106.14729, 2021.

[115] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the rgb-d slam system," in 2012 IEEE international conference on robotics and automation. IEEE, 2012, pp. 1691–1696.

[116] S. Stekovic, M. Rad, F. Fraundorfer, and V. Lepetit, "Montefloor: Extending mcts for reconstructing accurate large-scale floor plans," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16 034–16 043.

[117] Y. Yue, T. Kontogianni, K. Schindler, and F. Engelmann, "Connecting the dots: Floorplan reconstruction using two-level queries," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 845–854.

[118] M. Tabata, K. Kurata, and J. Tamamatsu, "Shape-net: Room layout estimation from panoramic images robust to occlusion using knowledge distillation with 3d shapes as additional inputs," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3551–3560.

[119] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras, "An integrated platform for live 3d human reconstruction and motion capturing," IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 4, pp. 798–813, 2017.

[120] C. Wu, X.-J. Wu, T. Xu, Z. Shen, and J. Kittler, "Motion complement and temporal multifocusing for skeleton-based action recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 1, pp. 34–45, 2024.