2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

A Collaborative Stereo Camera with Two UAVs for Long-distance Mapping of Urban Buildings

Zhaoying Wang, Wei Dong

***Abstract*— For a swarm of Unmanned Aerial Vehicle (UAV), long-distance visual mapping is advantageous for pre-planning navigation paths in unknown urban building environments. This work leverages two cameras on two UAVs to build a wide-baseline collaborative stereo camera, which can construct a navigable mesh map for remote building obstacles. We present a complete framework of the collaborative stereo camera for long-distance mapping, including online extrinsic parameter estimation of the stereo camera, real-time cross-camera feature association, and semantic mesh map generation of remote buildings. Extensive simulations and real-world experiments verify the effectiveness of the collaborative stereo camera. With a 3m baseline, the collaborative stereo camera achieves long-distance mapping of buildings (20m $\sim$ 50m) away with a relative error of approximately 10%. The constructed remote map enables UAVs to pre-detect large obstacles and pre-plan navigation paths in large-scale building environments. Hopefully, this work can provide a novel and practical approach for collaborative visual tasks of UAV swarm.**

Video - https://youtu.be/a0kj-1zb6KI

I. INTRODUCTION

Autonomous navigation of a swarm of Unmanned Aerial Vehicles (UAVs) [1] demands real-time mapping of obstacles. Compared with expensive and heavy laser lidar, the low-cost, lightweight stereo depth cameras are widely adopted in UAVs by providing depth point clouds to construct obstacle maps. Extensive research has successfully employed onboard stereo depth cameras (such as Intel Realsense D435) to construct real-time obstacle maps and enable UAVs to navigate in forests [2], indoors [3], and pedestrian [4] scenes. However, large-scale urban building obstacles may pose challenges. Large building obstacles typically span tens of meters [5]. While the existing stereo camera generally has an effective depth range of 10m, which is much smaller than the scale of building obstacles. When UAVs navigate in unknown building scenarios, UAVs only sense the depth information of the obstacles ahead when they fly close to the buildings. However, the huge building obstacles make it difficult for UAVs to find feasible navigation paths within their field of view. The Structure From Motion (SFM) of buildings by single UAV requires specifically planned UAV trajectory [6], which is not suitable for remote and unknown obstacles ahead. Extending the baseline of the stereo camera with two UAVs is a promising approach for long-distance mapping unknown obstacles ahead. The study in [7] simulates two wide-baseline downward-facing cameras carried by two UAVs in high altitude. However, due to wind interference and control errors in the real world, multiple UAVs may not provide the extremely stable binocular vision setup for the above wide-baseline stereo system.

The authors are with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wangzhaoying@sjtu.edu.cn; dr.dongwei@sjtu.edu.cn;).

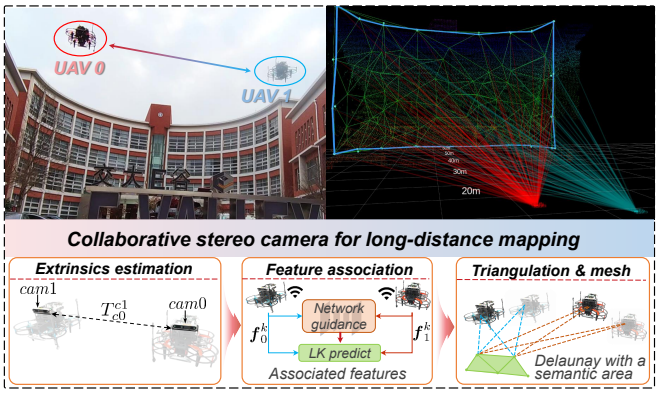

Fig. 1. The figure shows the real-world long-distance mapping experiment and the complete pipeline of the collaborative camera.

In this work, we propose a practical and complete framework for long-distance mapping of buildings by utilizing two UAVs as a collaborative stereo camera. First, we propose an online extrinsic parameter estimation for the stereo camera. Relative position is estimated by optimizing observations from visual markers, Inertial-Measurement-Unit (IMU), and Ultra-Wideband (UWB). In addition, a bidirectional visual observation algorithm is developed for accurate relative attitude estimation. Second, for the dynamic camera views of UAVs, we propose a dual-channel cross-camera feature association algorithm for two UAVs. The algorithm builds a parallel guidance channel and prediction channel, which achieves both accurate and real-time performance. Third, A long-distance mesh map is established based on landmarks and with a semantic expansion algorithm. The brief pipeline is presented in Fig.1. The hardware of the stereo camera is shown in Fig.2. We conclude the main contributions as follows:

- An online extrinsic parameter estimation algorithm is proposed for collaborative stereo camera in dynamic flight. Specifically, a fiducial Marker-based Visual-Inertial-Ranging Estimator (MVIRE) is designed for the estimation of the relative position between two cameras. A Bidirectional Visual Observation (BVO) algorithm is developed to calculate the relative attitude.
- A cross-camera feature association algorithm is pro-

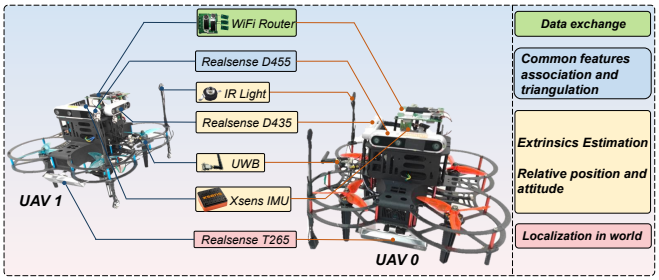

Fig. 2. The hardware setup of the collaborative stereo camera.

posed with a dual-channel structure, which achieves both accurate and real-time performance.

- A semantic mesh mapping algorithm for remote buildings is introduced, consisting of collaborative triangulation of landmarks, semantic area expansion, and Delaunay triangular generation.

II. Related Work

A. Extrinsic Parameter Estimation of Stereo Camera

The offline extrinsic parameter estimation of a fixed-baseline stereo camera is generally completed by using a checkerboard or Aprilgrids [8]. Since the two lenses in our collaborative camera are placed on two independent UAVs, we need to online estimate the varying camera extrinsic parameters in real-time. This problem can be converted to relative pose estimation between two UAVs. The study in [9] analyzes the visual fiducial markers, like ARTag, AprilTag, and ArUco, for relative pose estimation. The tag board facilitates quick deployment, but tag pattern detection is vulnerable to varying illumination and shadows in the outdoors. The active fiducial markers like infrared lights [10] can be used for robustly relative pose calculation under outdoor illumination with the Perspective-n-Points (PnP) algorithm. As the accuracy of marker-based visual estimation deteriorates with increasing observation distance, other sensors such as IMU and UWB are fused into the pose estimation to improve accuracy [11], [12]. However, these works mostly focus on the estimation of position rather than attitude. The relative attitude has a significant impact on the accuracy of collaborative mapping and is worthy of further investigation.

B. Cross-camera Feature Association

Cross-camera feature association between two UAVs requires highly precise matching performance. In addition, the real-time capability of feature association is necessary during the dynamic flight. The study in [13] thoroughly evaluates the well-known classic and learned feature association algorithms. The compared feature descriptors include ORB, SIFT, SURF and SuperPoint [14]. The compared matching method includes NN matcher and various outlier rejectors: RANSAC [15], and learned matcher SuperGlue [13]. The combination of SuperPoint with SuperGlue outperforms in precision compared with other approaches. The real-time performance of above detectors and matchors are compared in [16]. Unfortunately, the SuperPoint and SuperGlue impose a heavy computational burden on the resource-constrained onboard computer of UAV. The high-rate execution of SuperPoint and SuperGlue in real-time is infeasible for the onboard computer currently. Improving the real-time performance of the highly precise feature association is still worth further investigation.

C. Long-distance Mesh Mapping

Structure from motion (SFM) leverages a sequence of images from multiple viewpoints to construct a map of distant buildings [6]. In order to obtain an accurate map, the camera is expected to accumulate enough parallax in specific directions (such as parallel to the building surfaces). Similarly, single-camera SLAM also requires long-distance movement to triangulate distant map points in urban buildings. Adding multiple collaborative cameras in space can compensate for the long-distance movement of a single camera, saving time and improving real-time performance. The study in [7] uses two cooperative wide-baseline cameras to triangle 3D landmarks, but only in simulated environments with uniformly rich textures. However, the distribution of features on the building surface is uneven, with more features near windows and fewer features on walls. Each part of a building can pose a potential obstacle to UAVs. Thus, constructing building surfaces that are as complete as possible is worth investigating. Additionally, sparse 3D landmarks are not friendly for navigation. Kimera [17] and [5] utilize the triangled 3D landmarks to generate navigable mesh maps, but it is only effective for nearby feature points. Generating long-distance mesh maps covering large-scale buildings remains to be studied.

III. Extrinsic Parameter Estimation of Stereo Camera

A. Method Setup

Since each lens of the collaborative stereo camera is fixed on each UAV with the same layout, the online estimation of extrinsic parameters between two lenses equals to relative pose estimation between two UAVs. The relative pose estimation is divided into relative position estimation and relative attitude estimation for illustration.

B. Relative Position Estimation

We define camera C_0 on UAV0 as the primary camera, calculating the relative position of camera C_1 on UAV1. For an accurate estimation with multi-sensor fusion, the argument relative state $\mathcal{X}_{01}^k$ is established, including relative position $\boldsymbol{p}_{01}^k$ and velocity $\boldsymbol{v}_{01}^k$ at timestamp k. The state estimation is defined as $\hat{\mathcal{X}}_{01}^k = (\hat{\boldsymbol{p}}_{01}^k, \hat{\boldsymbol{v}}_{01}^k)$. We construct a sliding window of states as follows:

$$\hat{\mathcal{X}}_{01} = (\hat{\mathcal{X}}_{01}^{k-m}, \hat{\mathcal{X}}_{01}^{k-m+1}, ..., \hat{\mathcal{X}}_{01}^{k}) \tag{1}$$

where $m \in \mathbb{N}$ is the size of sliding window. To achieve an accurate estimation, we propose a fiducial Marker-based Visual-Inertial-Ranging Estimator (MVIRE) considering the observations from the following three factors:

- The Perspective-n-Point (PnP) observation of fiducial IR markers on UAV1 from the side camera (D435) of UAV0.
- The acceleration $\boldsymbol{a}_0$ from IMU0 on UAV0 and $\boldsymbol{a}_1$ from IMU1 on UAV1.
- The ranging measurement from two UWBs, which are placed on the two UAVs, respectively.

The cost function $f(\hat{\mathcal{X}}_{01})$ constructed from the above three factors is formulated as follows:

$$\begin{aligned} f(\hat{\mathcal{X}}_{01}) \triangleq \min_{\mathcal{X}_{01}} \Bigg\{ & \sum_{i=k-m}^{k} \left\| \mathbf{r}_{\mathcal{V}} \left(\mathbf{z}^i_{\mathcal{V}_{01}}, \hat{\mathcal{X}}_{01} \right) \right\|^2_{\mathbf{P}^i_{\mathcal{V}_{01}}} \\ & + \sum_{i=k-m}^{k} \left\| \mathbf{r}_{\mathcal{I}} \left(\mathbf{z}^i_{\mathcal{I}_0}, \mathbf{z}^i_{\mathcal{I}_1}, \hat{\mathcal{X}}_{01} \right) \right\|^2_{\mathbf{P}^i_{\mathcal{I}_{01}}} \\ & + \sum_{i=k-m}^{k} \left\| \mathbf{r}_{\mathcal{R}} \left(\mathbf{z}^i_{\mathcal{R}_{01}}, \hat{\mathcal{X}}_{01} \right) \right\|^2_{\mathbf{P}^i_{\mathcal{R}_{01}}} \Bigg\} \end{aligned} \tag{2}$$

where $\mathbf{r}_{\mathcal{V}}(\cdot), \mathbf{r}_{\mathcal{I}}(\cdot), \mathbf{r}_{\mathcal{R}}(\cdot)$ are the residuals constructed from Visual PnP, IMU, and UWB observations. $\mathbf{P}^i_{\mathcal{V}_{01}}, \mathbf{P}^i_{\mathcal{I}_{01}}, \mathbf{P}^i_{\mathcal{R}_{01}}$ are the covariance matrix of measurement noise.

The residual of the visual PnP factor is calculated as follows:

$$\mathbf{r}_{\mathcal{V}} \left(\mathbf{z}^i_{\mathcal{V}_{01}}, \hat{\mathcal{X}}_{01} \right) = \begin{bmatrix} \tilde{\boldsymbol{p}}^i_{\mathcal{V}_{01}} - \hat{\boldsymbol{p}}^i_{01} \\ \mathbf{0} \end{bmatrix} \tag{3}$$

where $\tilde{\boldsymbol{p}}^i_{\mathcal{V}_{01}}$ is relative position of UAV1 in UAV0 solved by PnP. The residual of IMU factor is presented as follows:

$$\mathbf{r}_{\mathcal{I}} \left(\mathbf{z}^i_{\mathcal{I}_0}, \mathbf{z}^i_{\mathcal{I}_1}, \hat{\mathcal{X}}_{01} \right) = \begin{bmatrix} \boldsymbol{R}^i_{01} \tilde{\boldsymbol{p}}^i_1 - \tilde{\boldsymbol{p}}^i_0 - \hat{\boldsymbol{p}}^i_{01} \\ \boldsymbol{R}^i_{01} \tilde{\boldsymbol{v}}^i_1 - \tilde{\boldsymbol{v}}^i_0 - \hat{\boldsymbol{v}}^i_{01} \end{bmatrix} \tag{4}$$

where $\tilde{\boldsymbol{p}}^i_0$ is the predicted position of UAV0 with acceleration from IMU0 as $\tilde{\boldsymbol{p}}^i_0 = \boldsymbol{p}^{i-1}_0 + \boldsymbol{v}^{i-1}_0 \Delta t + \frac{1}{2} \boldsymbol{a}^i_0 \Delta t^2$. The $\boldsymbol{p}_0, \boldsymbol{v}_0$ is w.r.t. the origin coordinate of UAV0. For velocity term, $\tilde{\boldsymbol{v}}^i_0$ is the predicted velocity of UAV0 with acceleration from IMU0 as $\tilde{\boldsymbol{v}}^i_0 = \boldsymbol{v}^{i-1}_0 + \boldsymbol{a}^i_0 \Delta t$. While $\tilde{\boldsymbol{p}}^i_1, \tilde{\boldsymbol{v}}^i_1$ represents the predicted position and velocity of UAV1 with the acceleration of IMU1 w.r.t. the origin coordinate of UAV1. $\boldsymbol{R}^i_{01}$ is to transfer the coordinate of UAV1 to UAV0. We note that the $\boldsymbol{a}_0, \boldsymbol{a}_1$ are transferred with the attitude of UAV and represented w.r.t. the origin coordinate of UAV0 and UAV1 respectively. The acceleration bias is calibrated before the flight in this work and will be added to online estimation in the future.

The residual of the UWB factor is illustrated as follows:

$$\mathbf{r}_{\mathcal{R}} \left(\mathbf{z}^i_{\mathcal{R}_{01}}, \hat{\mathcal{X}}_{01} \right) = \begin{bmatrix} d^i_{\mathcal{R}_{01}} - \left\| \hat{\boldsymbol{p}}^i_{01} \right\| \\ \mathbf{0} \end{bmatrix} \tag{5}$$

where $d^i_{\mathcal{R}_{01}}$ is the distance measured by UWB between UAV0 and UAV1. Finally, the relative state $\mathcal{X}^k_{01}$ can be efficiently solved by our dimension-reduced wriggling estimator [12]. In addition, the PnP calculation on the fiducial IR markers adopts a robust dual-source positioning algorithm, which is introduced in our previous work [10].

C. Relative Attitude Estimation

Due to the alignment of gravity, the roll and pitch angles $(\theta_0, \phi_0, \theta_1, \phi_1)$ measured by the IMU0 of UAV0 and IMU1 of UAV1 typically exhibit high precision and consistency in the global coordinate system. Thus the relative roll and pitch can be directly calculated as $\theta_{01} = \theta_1 - \theta_0$ and $\phi_{01} = \phi_1 - \phi_0$ respectively. However, affected by the disturbance of the environmental magnetic field, the yaw angles (ψ_0, ψ_1) among two UAVs often show inconsistency in the global coordinate system. Consequently, we propose a bidirectional visual observation (BVO) method to compute the relative yaw ψ_{01} between UAVs.

As shown in Fig.3(a), the IR camera of Realsense D435 and the center IR light marker are placed closely and vertically aligned on the side of the UAV. Thus, we can regard the location of the IR marker as the location of the IR camera in the horizontal plane or in the bird's view. As shown in Fig.3(b), the IR marker on UAV1 is observed by the IR camera of D435 on UAV0 with a view angle α_0. Meanwhile, the IR marker on the UAV0 is observed by the IR camera of D435 on UAV1 with a view angle α_1. Since the observed IR marker could represent the IR camera in the horizontal plane, α_0, α_1 are also the respective observation angles for two cameras mutually observing each other. Clearly, the relative yaw of UAVs can be calculated by the difference of observation angles as $\psi_{01} = \alpha_1 - \alpha_0$. When the attitudes of two UAVs are horizontal (roll and pitch angles are nearly zero), the calculation details of α_0 and α_1 are shown in Fig.3(c) and Fig.3(d) respectively. We take α_0 as an example. The pixel location of the center IR marker is (u_0, v_0). The view angle α_0 can be calculated with the IR pixel location in the x-axis by $\alpha_0 = (u_0 - c_{x0})/f_{x0}$, where (f_{x0}, c_{x0}) are the focal length and principal point in the x-axis. A similar calculation of view angle α_1 is illustrated in Fig.3(d). Finally, the relative yaw ψ_{01} can be calculated as follows:

$$\psi_{01} = (u_1 - c_{x1})/f_{x1} - (u_0 - c_{x0})/f_{x0} \tag{6}$$

Then we consider the pixel location of the IR marker affected by the roll and pitch attitude of the UAV. We take UAV0 as an example. Roll θ_0 affects the pixel location of IR marker along the y-axis of the image as $v^\theta_0 = v_0 - (\theta_0 f_{y0} + c_{y0})$. Pitch ϕ_0 rotates the pixel location (u_0, v_0) around the optical center as $(u^\phi_0, v^\phi_0)^T = R_{\phi_0}(u_0, v_0)^T$. Thus we can utilize θ_0, ϕ_0 to calculate the origin IR pixel location when UAV0 is at horizontal attitude. A similar calculation can be completed for UAV1. Then, we can use Eq.6 to calculate relative yaw.

IV. Cross-camera Feature Assocaition

This part introduces real-time common feature association in the overlapping view of two camera perspectives. The feature association should guarantee both high accuracy and real-time performance. The dual-channel cross-camera feature association method from our previous work [18] is adopted in this part. As shown in Fig.4, the images captured

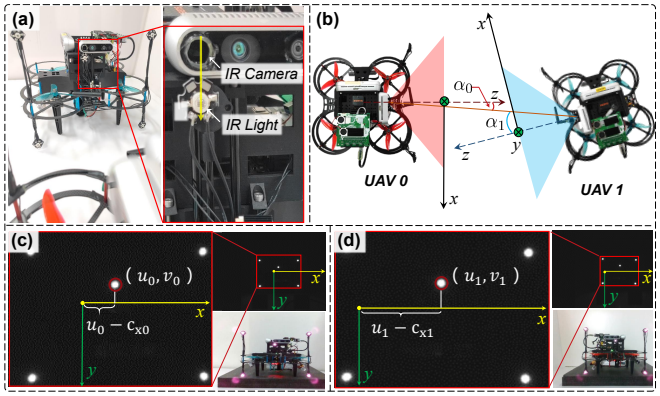

Fig. 3. The illustration of bidirectional visual observation.

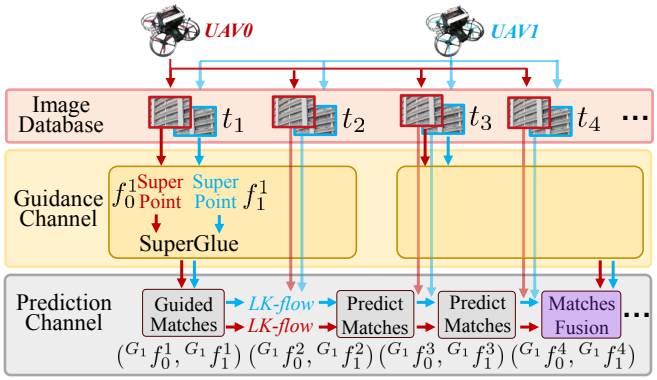

Fig. 4. The schema of dual-channel feature association for two UAVs.

by UAV0 and UAV1 constitute the paired images in image database. A parallel guidance channel and prediction channel structure are designed. The guidance channel is responsible for extracting SuperPoint features (f_0, f_1) from the paired images of two UAVs and matching the features with SuperGlue. Generally, the SuperPoint and SuperGlue can produce highly accurate feature matches. We call associated features as guided feature matches $({}^G f_0, {}^G f_1)$. However, the execution of SuperPoint and SuperGlue on the onboard computer is time-consuming. In order to improve the real-time performance, the prediction channel receives $({}^G f_0, {}^G f_1)$ from the guidance channel, and then adopts Lucas-Kanade optical flow (LK-flow) to predict and track the features $({}^G f_0, {}^G f_1)$ with new image pairs. The features are tracked from 1 to k frames as $({}^G f_0^1, {}^G f_1^1, {}^G f_0^2, {}^G f_1^2, ..., {}^G f_0^k, {}^G f_1^k)$ until feature moves out of any side camera image. During the optical flow tracking process, the correspondence between feature points ${}^G f_0$ and ${}^G f_1$ is retained. Consequently, we can obtain the new feature associations with new image pairs in real-time.

As the camera field of view shifts, the tracked feature points gradually move out of the current perspective, resulting in a continuous reduction of common features. Thus, the guidance channel periodically provides newly associated features for the prediction channel at a low rate. Then the new features and existing features in the prediction channel are fused [18]. In conclusion, the novel design of periodic guidance of SuperPoint and SuperGlue and fast prediction with LK-flow effectively achieves a highly accurate and real-time feature association between two UAVs.

V. Long-distance Mesh Mapping

Long-distance mesh mapping is based on accurate extrinsic parameter estimation of stereo camera and common feature association. In this module, 3D landmarks are collaboratively triangulated from 2D associated features with the accurate relative camera pose. Then a semantic mesh map is constructed based on these landmarks.

A. Collaborative Triangulation

This part introduces the collaborative triangulation of landmarks through the associated features. An accurate and robust landmark triangulation requires sufficient observations of features. Suppose the landmark $\mathbf{p}_f$ is observed by the camera C_0 of UAV0 and C_1 of UAV1 from 1 to k timestamp. The observation of landmark $\mathbf{p}_f$ in C_0 and C_1 at k timestamp are represented as ${}^{C_0^k}\mathbf{p}_f, {}^{C_1^k}\mathbf{p}_f$. For the collaborative triangulation, we select the camera frame of UAV0 at k timestamp as the anchor frame ${}^A(\cdot)$. The $\mathbf{p}_f$ in the anchor frame is also denoted as ${}^A\mathbf{p}_f$. Then we leverage other feature observations of both UAV0 and UAV1 from 1 to k timestamp into this anchor frame by the following equation:

$$ {}^A\mathbf{p}_f = {}^A\mathbf{R}_{C_i^m}\, {}^{C_i^m}\mathbf{p}_f + {}^A\mathbf{p}_{C_i^m}, i = 0, 1, m = 1, 2, ..., k \quad (7) $$

where the camera pose of UAV0 and UAV1 w.r.t. anchor frame are denoted as ${}^A\mathbf{p}_{C_0^m}, {}^A\mathbf{R}_{C_0^m}, {}^A\mathbf{p}_{C_1^m}, {}^A\mathbf{R}_{C_1^m}$, respectively. The ${}^{C_i^m}\mathbf{p}_f$ can be further formulated as the multiplication of depth ${}^{C_i^m}z_f$ and bearing vector ${}^{C_i^m}\mathbf{b}_f$.

$$ \begin{aligned} {}^A\mathbf{p}_f &= {}^A\mathbf{R}_{C_i^m}\, {}^{C_i^m}z_f\, {}^{C_i^m}\mathbf{b}_f + {}^A\mathbf{p}_{C_i^m} \\ &= {}^{C_i^m}z_f\, {}^A\mathbf{b}_{C_i^m \to f} + {}^A\mathbf{p}_{C_i^m} \end{aligned} \quad (8) $$

Then we remove the degree of freedom of depth ${}^{C_i^m}z_f$ by multiplying the orthogonal space ${}^A\mathbf{N}_i^m = \lfloor {}^A\mathbf{b}_{C_i^m \to f} \times \rfloor$ to the bearing vector. We can get the following equation:

$$ \begin{aligned} {}^A\mathbf{N}_i^m\, {}^A\mathbf{p}_f &= {}^A\mathbf{N}_i^m {}^{C_i^m}z_f\, {}^A\mathbf{b}_{C_i^m \to f} + {}^A\mathbf{N}_i^m\, {}^A\mathbf{p}_{C_i^m} \\ &= {}^A\mathbf{N}_i^m\, {}^A\mathbf{p}_{C_i^m} \end{aligned} \quad (9) $$

We stack all the measurements of UAV0 and UAV1 from 1 to k timestamp.

$$ \underbrace{\begin{bmatrix} {}^A\mathbf{N}_0^1 \\ \vdots \\ {}^A\mathbf{N}_0^k \\ {}^A\mathbf{N}_1^1 \\ \vdots \\ {}^A\mathbf{N}_1^k \end{bmatrix}}_{\mathbf{A}_{6m\times 3}} {}^A\mathbf{p}_f = \underbrace{\begin{bmatrix} {}^A\mathbf{N}_0^1\, {}^A\mathbf{p}_{C_0^1} \\ \vdots \\ {}^A\mathbf{N}_0^k\, {}^A\mathbf{p}_{C_0^k} \\ {}^A\mathbf{N}_1^1\, {}^A\mathbf{p}_{C_1^1} \\ \vdots \\ {}^A\mathbf{N}_1^k\, {}^A\mathbf{p}_{C_1^k} \end{bmatrix}}_{\mathbf{b}_{6m\times 1}} \quad (10) $$

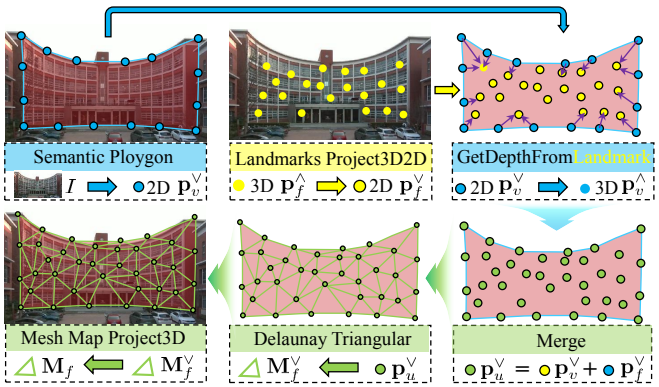

Fig. 5. The pipeline of Delaunay mesh map with semantic area expansion.

For the quick solution of ${}^{A}\mathbf{p}_f$, we reformulate the $6m \times 3$ system to 3×3 system by the following equation.

$$\mathbf{A}^{\top}_{6m\times 3}\mathbf{A}_{6m\times 3}{}^{A}\mathbf{p}_f = \mathbf{A}^{\top}_{6m\times 3}\mathbf{b}_{6m\times 1} \tag{11}$$

Finally, we additionally check the solved ${}^{A}\mathbf{p}_f$ by evaluating the condition number. The ${}^{A}\mathbf{p}_f$ with condition number over the pre-defined threshold will be removed.

B. Mesh Map Generation

For simplicity, we may omit the subscript ${}^{A}(\cdot)$ for the anchor frame when there is no ambiguity. We also add subscript $(\cdot)^{\vee}$ for variable in 2D image and $(\cdot)^{\wedge}$ for variable in 3D space. A traditional pipeline of mesh map generation based on landmarks is as follows. The 3D landmarks $\mathbf{p}_f^{\wedge}$ are first projected onto the image plane to obtain 2D vertices $\mathbf{p}_f^{\vee}$. $\mathbf{p}_f^{\vee}$ are then used to construct Delaunay mesh triangular grid maps $\mathbf{M}_f^{\vee}$ [5]. Finally, through associating the corresponding 3D landmarks of 2D vertices, the triangular grid maps $\mathbf{M}_f^{\vee}$ are projected back 3D space to construct the mesh map $\mathbf{M}_f^{\wedge}$.

The limitation of the existing pipeline lies in the fact that the mesh map can only cover areas where landmarks are present. When some building areas are textureless, these areas cannot generate enough landmarks for mesh grid maps. However, for the safe navigation of a UAV swarm, all building parts in the view should be represented as comprehensively as possible by the mesh map.

To address this issue, we propose a semantic area expansion method to expand the mesh map to these feature-missing areas as illustrated in Fig.5. This process involves the following steps. 1) Select the building area in the 2D image I with semantic segmentation (DeepLabv3+ [19]). 2) Fit a polygon P to the semantic area and extract the polygon vertices $\mathbf{p}_v^{\vee}$. 3) For each vertex, find the nearest $\mathbf{p}_f^{\vee}$ and assign the depth value of $\mathbf{p}_f^{\wedge}$ to that vertex. 4) Merge the feature points $\mathbf{p}_v^{\vee}$ along with those $\mathbf{p}_f^{\vee}$ as universal set $\mathbf{p}_u^{\vee}$. 5) Construct Delaunay triangular $\mathbf{M}_f^{\vee}$ from $\mathbf{p}_u^{\vee}$ and project back to 3D space as $\mathbf{M}_f^{\wedge}$. The whole process is concluded in the algorithm 1.

Algorithm 1 Mesh Map with Semantic Area Expansion

Input: Image I and pose $\mathbf{t}_0, \mathbf{q}_0$ of UAV0, landmark $\mathbf{p}_f^{\wedge}$.
Output: The mesh map $\mathbf{M}_f^{\wedge}$

P = **semantic**(I)
$\mathbf{p}_v^{\vee}$ = **findPloygonVertice**(P)
$\mathbf{p}_f^{\vee}$ = **project3D2D** $(\mathbf{p}_f^{\wedge}, \mathbf{t}_0, \mathbf{q}_0)$
for $\mathbf{p}_{vi}^{\vee}$ in $\mathbf{p}_v^{\vee}$ **do**
  for $\mathbf{p}_{fj}^{\vee}$ in $\mathbf{p}_f^{\vee}$ **do**
    d_j = **eulerDistance**$(\mathbf{p}_{fj}^{\vee}, \mathbf{p}_{vi}^{\vee})$
    $j = j + 1$
  end for
  $\mathbf{p}_{f\star}^{\vee} \leftarrow$ **minDistance**(d_j)
  $d_{\star}$ = **getDepth**$(\mathbf{p}_{f\star}^{\wedge})$
  $\mathbf{p}_{vi}^{\wedge}$ = **project2D3D**$(\mathbf{p}_{vi}^{\vee}, d_{\star}, \mathbf{t}_0, \mathbf{q}_0)$
  $i = i + 1$
end for
$\mathbf{p}_u^{\vee}$ = **merge**$(\mathbf{p}_v^{\vee}, \mathbf{p}_f^{\vee})$
$\mathbf{M}_f^{\vee}$ = **delaunayTriangular**$(\mathbf{p}_u^{\vee})$
$\mathbf{M}_f^{\wedge}$ = **project3D**$(\mathbf{M}_f^{\vee}, \mathbf{p}_v^{\wedge}, \mathbf{p}_f^{\wedge})$

VI. Experiments

A. Experiment setup

Experiments are performed with the two custom-developed quadcopter UAVs. The sensor layout of the UAV is shown in Fig.6. Each UAV is equipped with an Intel Realsense D455 as the forward-facing camera, of which the color lens is used as one of the lenses of the proposed collaborative stereo camera. The Intel Realsense D435 is mounted on the side of the UAV for mutual observation of fiducial IR markers. The fiducial IR markers consist of 5 infrared lights (850nm wavelength). Four lights (IR_1, IR_2, IR_3, IR_4) are arranged at the periphery to form the vertices of a square outline, facilitating the PnP solution for relative position estimation between two UAVs. One central light IR_5 is positioned directly beneath the D435 infrared lens, serving to indicate the position of the infrared lens and aiding in relative yaw angle calculation. The Nooploop UWB radio (model LinkTrack LTPS) is placed on the side of the UAV for range measurement between two UAVs. The IMU (model Xsens MTi-630) is mounted on the center of the UAV for acceleration measurement and attitude measurement. The Intel Realsense T265 is mounted at a 45-degree downward angle in front of the UAV, providing position in the world coordinate and facilitating feedback control for autonomous flight. WiFi mesh router (model TP-Link WDR7650) is placed on the top of the UAV to construct a local mesh network between two UAVs, supporting high-bandwidth and low-latency data exchange. An NVIDIA Jetson NX Xavier is adopted as the onboard computer.

B. Evaluation of Extrinsic Parameter Estimation

The evaluation of the extrinsic parameter estimation of the stereo camera is performed under the NOKOV motion capture system as shown in Fig.7. Both UAVs are manually controlled by operators to fly in random trajectories. The

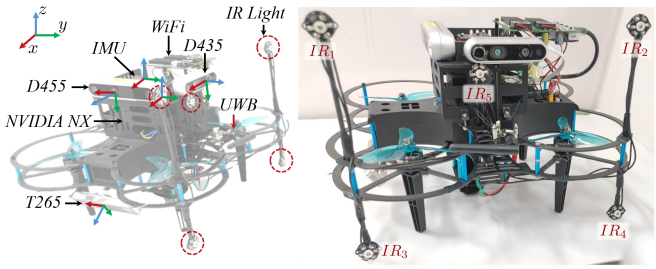

Fig. 6. The sensors layout of the custom-developed quadcopter UAV.

Fig. 7. Experiments of extrinsic parameter estimation of the stereo camera.

relative position and attitude of the two UAVs are estimated by our proposed MVIRE and BVO, respectively. A total of 3 sets of experiments $\mathcal{D}_1 \sim \mathcal{D}_3$ are performed, and the error analysis of position and orientation is recorded in Table I and Table II. The estimation of relative position and relative attitude of $\mathcal{D}_1$ are plotted in Fig.8 with the ground truth as comparisons. The RMSE of position by MVIRE is about 0.06m on the x-axis, 0.05m on the y-axis and 0.03m on the z-axis. The RMSE of relative roll and pitch (ϕ, θ) error is about 1.8° by IMU measurement difference and relative yaw ψ error is under 0.3° by the BVO algorithm.

To validate the effective range and accuracy of relative yaw ψ estimation, we conduct experiments $\mathcal{D}_4, \mathcal{D}_5$ by rotating two UAVs on a horizontally rotating platform. The relative yaw angle is estimated by BVO. The experiments are conducted in two sets with baseline distances d_{c0}^{c1} of 1.4m and 2.0m between the two UAVs, respectively. The estimated yaw ψ with ground truth is plotted in Fig.9. The statistics are analyzed in Table III. The effective range for relative yaw angles is from -30° to 30° with a baseline of 2.0m. The RMSE is under 0.36°.

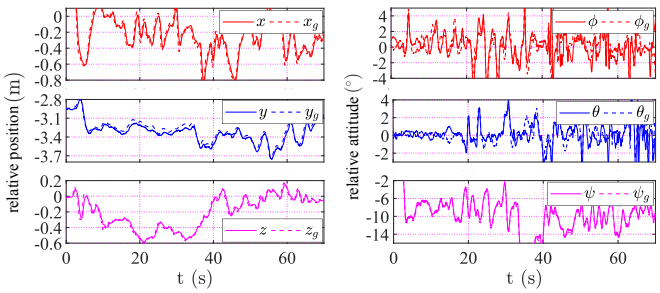

Fig. 8. The estimated extrinsic parameter of the stereo cameras with motion capture as ground truth. The orientation is represented by roll ϕ, pitch θ, yaw ψ. The ground truth is denoted as $(\cdot)_g$.

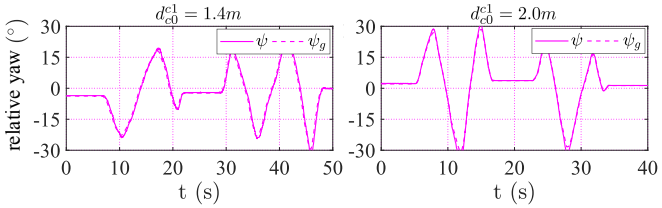

Fig. 9. The relative yaw estimation with the baseline of 1.4m and 2.0m.

TABLE I
RELATIVE POSITION ERROR ANALYSIS

Data Set	RMSE (m)			STD (m)		
	x	y	z	x	y	z
$\mathcal{D}_1$	0.0601	0.0505	0.0320	0.0600	0.0303	0.0320
$\mathcal{D}_2$	0.0595	0.0510	0.0314	0.0578	0.0298	0.0330
$\mathcal{D}_3$	0.0610	0.0486	0.0308	0.0606	0.0281	0.0301

C. Cross-camera Feature Association

We select four buildings on the campus to conduct feature association experiments. The image stream of each camera is configured as 640 × 480 pixels at 30 Hz. The matching results are shown in Fig.10. We also do run-time statistics in these four scenarios. When 200 key points are extracted for each UAV, the average run-time of the dual-channel algorithm is 14.1ms. The minimum and maximum run-time are 7.5ms and 26.4ms, respectively. The detailed run-time analysis can be found in our previous work [18].

D. Accuracy Analysis of Mesh Map

We evaluate the accuracy of the mesh map under different camera baselines and distances to buildings. We also explore the impact of pose errors on mapping accuracy. The experiments are executed in the simulated GAZEBO world. Two Iris UAVs are equipped with D455 cameras and employed with PX4 SITL for flight control. The GAZEBO scenario and the mesh map are shown in Fig.11. Two UAVs fly towards the building at a speed of 1m/s. The window length for landmark triangulation is set to 10 frames with a time interval of 0.05s for each UAV, comprising a total of 20 frames for the collaborative stereo camera.

Paired experiments are conducted for different camera

TABLE II
RELATIVE ORIENTATION ERROR ANALYSIS

Data Set	RMSE (°)			STD (°)		
	ϕ	θ	ψ	ϕ	θ	ψ
$\mathcal{D}_1$	1.8278	1.8385	0.2947	1.8037	1.8174	0.2942
$\mathcal{D}_2$	1.8054	1.7731	0.3051	1.7852	1.7516	0.3002
$\mathcal{D}_3$	1.7955	1.8044	0.2925	1.7921	1.8001	0.2887

TABLE III
RELATIVE YAW ERROR ANALYSIS

DataSet	$\mathbf{d}_{c0}^{c1}$ (m)	Min(°)	Max(°)	RMSE (°)	STD (°)
$\mathcal{D}_4$	1.4	-30	20	0.3242	0.2974
$\mathcal{D}_5$	2.0	-30	30	0.3571	0.3141

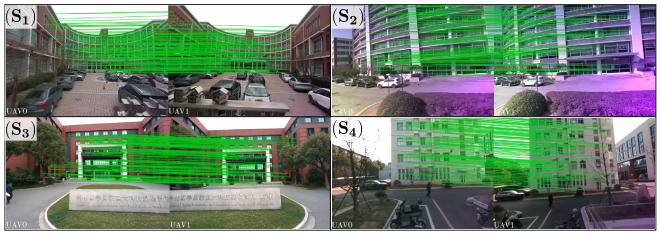

Fig. 10. The screenshot of the cross-camera feature association by collaborative UAVs in four building scenarios.

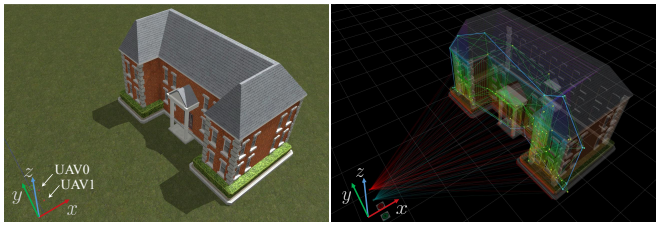

Fig. 11. (a) shows the collaborative UAVs constructing a mesh map of a school building in the GAZEBO world. (b) presents the obtained mesh map of the front surface of the building.

baselines d_{c0}^{c1} and varying distances d_{c0}^{b} to buildings. The baseline between the two UAVs is configured from 1m to 4m at intervals of 0.5m. The building is placed in front of the cameras at intervals of 5m, ranging from 20m to 50m. The ground truth of the building surface is obtained by the depth camera on the Iris UAV, whose valid depth range is set to 100m. The Euclidean distance between triangulated landmarks and their nearest ground truth map points is considered as the error ϵ. The mean error of all map points is denoted as ϵ_m. The error statistics are plotted in Fig.12. The experiments indicate that map error decreases with increasing the camera baseline and increases with longer distances to buildings.

Additionally, we investigate the influence of the extrinsic parameter errors of the stereo camera on map errors. The coordinate system adopts the front-left-up (FLU) rule. The camera baseline is along the y axis. For a baseline of 3m and a building distance of 40m, we set the basic relative position and attitude of UAV1 from UAV0 as (0m, 3m, 0m) and (0°, 0°, 0°) respectively. Then we add pose error biases for the

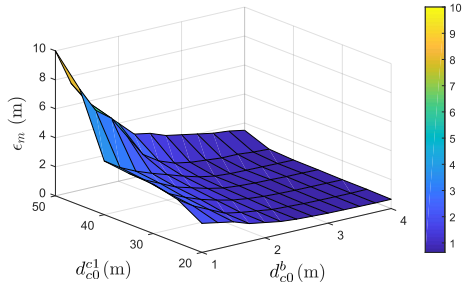

Fig. 12. The mean error ϵ_m of the mesh map under different baselines d_{c0}^{c1} and building distances d_{c0}^{b}.

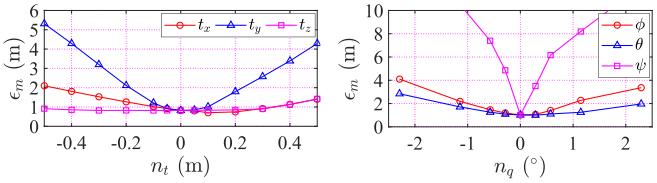

Fig. 13. The map error under relative position and attitude error bias.

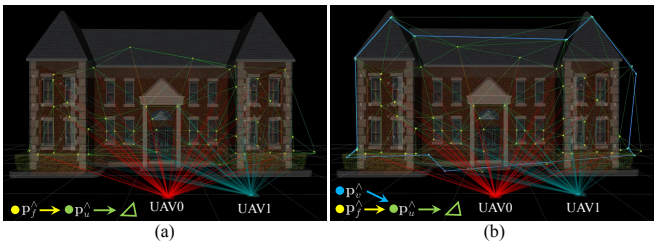

Fig. 14. The Delaunay mesh map of the school building. (a) shows the mesh map only using the triangled landmarks. (b) shows the mesh map with the addition of building edge vertices by semantic area expansion.

stereo camera. As shown in Fig.13, the relative position error bias is set from -0.5m to 0.5m. The relative position error along the baseline direction has the greatest impact on map errors. The relative attitude error in yaw ψ has the most significant impact on map errors, with a 1° error resulting in approximately 8m of map error. That is why we propose the BVO algorithm to pay more attention to relative yaw estimation.

E. Semantic Delaunay Mesh Map

We evaluate the method of semantic area expansion for Delaunay mesh maps. In Fig.14 (a), we triangulate 50 landmarks $\mathbf{p}_f^{\wedge}$ (highlighted in yellow) from the front surface of the building. Only $\mathbf{p}_f^{\wedge}$ constitute vertices $\mathbf{p}_u^{\wedge}$ to construct a Delaunay triangular mesh. It can be observed that the window areas of the building are covered with mesh map. However, the roof is not covered by the mesh map due to the missing features and landmarks. As depicted in Fig.14 (b), To address the feature-missing area, we employ semantic area expansion by obtaining 12 polygon vertices $\mathbf{p}_v^{\wedge}$ (highlighted in blue) from the semantic regions of the building. The landmarks $\mathbf{p}_f^{\wedge}$ and polygon vertices $\mathbf{p}_v^{\wedge}$ are both used to construct a Delaunay triangular mesh. The roof area is effectively supplemented with a Delaunay mesh map. Additionally, we calculate the coverage ratio of the mesh map on the building in the 2D image. The proposed semantic area expansion improves the coverage ratio of the mesh map from 70% to 92% in this scenario.

TABLE IV
MESH MAP ERROR ANALYSIS IN REAL-WORLD SCENARIOS

Scenario	d_{c0}^{c1} (m)	d_{c0}^{b} (m)	ϵ_m (m)	δ (%)
$\mathbf{S_1}$	2.8	51	4.53	8.88
$\mathbf{S_2}$	3.2	28	2.98	10.64
$\mathbf{S_3}$	3.0	43	3.92	9.11
$\mathbf{S_4}$	3.1	32	3.31	10.34

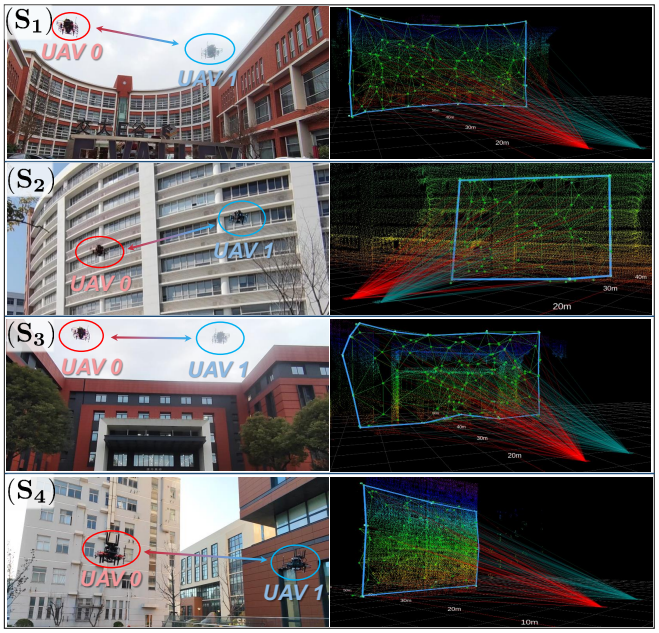

Fig. 15. The long-distance mesh mapping of urban buildings using collaborative UAVs in real-world scenarios.

F. Long-distance Mesh Mapping in Real-World

We conduct real-world mesh mapping experiments on urban buildings using our collaborative stereo camera by two UAVs. Four building scenarios and mesh map results are shown in Fig.15. Ground truth of building surfaces is obtained using the Livox Horizontal LiDAR, which is fixed on the ground. The LiDAR coordinate is aligned with the home coordinate system of the UAV. The mean map error ϵ_m with ground truth and relative map error percentage δ with the building distance in four scenarios are listed in Table IV. The relative map error δ is about $9\% \sim 11\%$ with a building distance from 20m to 50m. Although long-distance mapping may not be very accurate in the real world, this level of precision can support UAVs to pre-plan large-scale navigatable paths in advance.

VII. CONCLUSIONS

In this paper, a collaborative stereo camera using two UAVs is proposed for long-distance mapping of urban buildings. We first develop MVIRE and BVO algorithms to online estimate the extrinsic parameter of the stereo camera. Then a dual-channel structure is developed for the cross-camera feature association with both real-time and high-accuracy performance. Finally, collaborative landmark triangulation and semantic mesh expansion are performed to establish mesh maps for long-distance buildings. Real-world experiments show that the collaborative stereo camera with a 3m baseline achieves 20m $\sim$ 50m mapping of buildings with an average relative error of approximately 10%.

Currently, we depend more on two wide-baseline camera views for landmarks triangulation with a short time horizon. In the future, we will collect more observations with a longer time horizon and actively plan the camera view of each UAV to facilitate high-precise mapping. Furthermore, we plan to add more UAVs to form a larger aerial camera network for collaborative mapping tasks.

REFERENCES

[1] X. Zhou, X. Wen, Z. Wang, Y. Gao, H. Li, Q. Wang, T. Yang, H. Lu, Y. Cao, C. Xu, and F. Gao, "Swarm of micro flying robots in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm5954, 2022.

[2] A. Loquercio, E. Kaufmann, R. Ranftl, M. Müller, V. Koltun, and D. Scaramuzza, "Learning high-speed flight in the wild," *Science Robotics*, vol. 6, no. 59, p. eabg5810, 2021.

[3] G. Chen, W. Dong, X. Sheng, X. Zhu, and H. Ding, "An active sense and avoid system for flying robots in dynamic environments," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 2, pp. 668–678, 2021.

[4] G. Chen, P. Peng, P. Zhang, and W. Dong, "Risk-aware trajectory sampling for quadrotor obstacle avoidance in dynamic environments," *IEEE Transactions on Industrial Electronics*, 2023.

[5] L. Teixeira and M. Chli, "Real-time mesh-based scene estimation for aerial inspection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4863–4869.

[6] S. Daftry, C. Hoppe, and H. Bischof, "Building with drones: Accurate 3d facade reconstruction using mavs," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 3487–3494.

[7] M. Karrer and M. Chli, "Distributed variable-baseline stereo slam from two uavs," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 82–88.

[8] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 4304–4311.

[9] M. Kalaitzakis, B. Cain, S. Carroll, A. Ambrosi, C. Whitehead, and N. Vitzilaios, "Fiducial markers for pose estimation: Overview, applications and experimental comparison of the artag, apriltag, aruco and stag markers," *Journal of Intelligent & Robotic Systems*, vol. 101, no. 4, p. 71, 2021.

[10] Z. Wang, S. Liu, G. Chen, and W. Dong, "Robust Visual Positioning of the UAV for the Under Bridge Inspection With a Ground Guided Vehicle," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[11] H. Xu, Y. Zhang, B. Zhou, L. Wang, X. Yao, G. Meng, and S. Shen, "Omni-swarm: A decentralized omnidirectional visual–inertial–uwb state estimation system for aerial swarms," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3374–3394, 2022.

[12] W. Dong, Z. Mei, Y. Ying, S. Chen, X. Zhu *et al.*, "Sribo: An efficient and resilient single-range and inertia based odometry for flying robots," *arXiv preprint arXiv:2211.03093*, 2022.

[13] 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.

[14] 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.

[15] K. G. Derpanis, "Overview of the ransac algorithm," *Image Rochester NY*, vol. 4, no. 1, pp. 2–3, 2010.

[16] D. Bojanić, K. Bartol, T. Pribanić, T. Petković, Y. D. Donoso, and J. S. Mas, "On the comparison of classic and deep keypoint detector and descriptor methods," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2019, pp. 64–69.

[17] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, 2022.

[18] Z. Wang and W. Dong, "Real-time estimation of relative pose for uavs using a dual-channel feature association," *arXiv preprint arXiv:2402.17504*, 2024.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.