# Toward Cooperative 3D Object Reconstruction with Multi-agent

Xiong Li, Zhenyu Wen, Leiqiang Zhou, Chenwei Li, Yejian Zhou, Taotao Li and Zhen Hong*

***Abstract*— We study the problem of object reconstruction in a multi-agent collaboration scenario. Specifically, we focus on the reconstruction of specific goals through several cooperative agents equipped with vision sensors to achieve higher efficiency than single agents. Our main insight is that a complete 3D object can be split into several local 3D models and assigned to different agents. In addition, we can use the salient characteristics of the collaboration agent itself to help realize the integration of local models. We develop a novel pipeline that first restores local 3D models from the images obtained from different agents, then the relative poses between collaborative agents are estimated by aligning intrinsic features. After that, all local models are integrated using the estimated parameters. Extensive experiments show that our proposed method is capable of accurately reconstructing 3D objects in the real world in a multi-agent collaborative manner. The full reconstruction pipeline is released to the public as an open-source project.**

## I. INTRODUCTION

Robotics have been widely used in many areas such as search and rescue, infrastructure inspection, and exploration [12], [18], [25]. In recent years, researchers are focusing on improving the maneuverability of robot clusters, aiming to collaboratively perform operations such as reconnaissance and land surveying [26], [40]. To realize the above-mentioned operations, multi-agent 3D object reconstruction is one of the key challenging tasks. Fig. 1 illustrates the high-level view of the task. Each agent (camera) can only obtain partial visual information from the target object, and then the collected visual information is sent to a central server for 3D object reconstruction.

Previous studies [4], [7], [27], [45] have solved well the problem of object reconstruction in the case of continuous shooting by a single agent. However, multi-agent 3D object reconstruction brings the following challenges. First, it is difficult for multiple agents to achieve highly consistent scene understanding. In other words, they cannot accurately perceive the pose relationship between each other to correlate all visual information. Secondly, the scale error makes it difficult to correlate the relative pose relationships between agents with the corresponding viewpoints in the reconstructed space. In addition, some end-to-end solutions [22], [35], [38] suffer from the generalization issues that may be affected by light, and background noise which is naturally included in real-world applications.

Our key insight is that object reconstruction tasks based on multi-agent collaboration can be effectively decomposed. On the one hand, we can decompose a complete object into several local 3D models, which can be spliced. On the other hand, the information needed to infer the local model is already included in the images taken from each viewpoint.

We find that the integration of local 3D models can be achieved using only some iconic features of the cooperative robot itself. We describe it as a registration problem in point cloud space, which provides a stable and effective way to estimate the pose conversion relationship between agents. In order to create a local 3D model of the target object from the scene, we must also accurately eliminate the interference of background noise during the 3D mapping of the image. For example, even if an object has a complex geometric structure and a high degree of fit with the environment, it can be accurately extracted from the scene.

We propose a novel two-stage pipeline for reconstructing specific objects in a scene in the case of difficulties in continuous image acquisition. First, we generate local 3D models from all agents' perspectives with the help of a stereo-matching network. Secondly, we estimate the mutual pose relationships by matching the pre-added salient characteristics of the cooperative agents and use them to restore the complete 3D object. Our key contributions are:

1) To the best of our knowledge, we are the first to propose a multi-agent collaborative 3D reconstruction scene, which requires only a few images to achieve superior 3D reconstruction performance and has significant application prospects in the future.
2) We introduce a single-viewpoint object reconstruction method that enables each agent to reach a consensus on the target object, overcomes the challenge of environmental interference, and recovers the local 3D model of the target object at each viewpoint.
3) We develop a marker-based robust point cloud concatenation algorithm that leverages the cooperative relationship between agents to establish stable connections between widely varying viewpoints, and then quickly integrate all local information together with low computational cost.

## II. RELATED WORK

**Traditional 3D Reconstruction.** Traditional 3D reconstruction is usually divided into two stages, namely Structure from Motion (SfM) and Multi-View Stereo (MVS). The SfM [8], [13], [27] first estimates the camera motion and acquires a sparse point cloud model, after which MVS [16], [33], [43] is responsible for densifying the model to bring it closer to the real scene. Recently, neural methods for new view synthesis [17], [32] have also been proven to be able

The authors are with the Institute of Cyberspace Security and the College of Engineering, Zhejiang University of Technology.

*Corresponding author: Z. Hong

to replace MVS and achieve the high-quality dense reconstruction of objects. However, these methods require highly overlapping views and Lambertian surfaces to better realize feature extraction and tracking between different views. In addition, its application scenarios are further limited due to huge computing resources and time costs.

**End-to-end 3D Reconstruction.** Object reconstruction based on deep learning [9], [14] aims to avoid complicated camera calibration processes and realize end-to-end mapping from 2D images to 3D models. In this type of method, the input can be single [20], [36], [39] or multiple [5], [31], [38] RGB images, and the output has various forms such as voxel, point cloud, grid, etc. When the input is a video stream [6], [30], temporal correlation can be used to improve the smoothness and inter-frame consistency of the reconstruction. However, the above methods are usually only trained and evaluated on public datasets. On the one hand, they still do not have strong generalization capabilities for real-world applications. On the other hand, the large number of parameters included in the network model also leads to expensive computational costs.

In this work, we integrate different views with a low overlap rate to achieve the complete reconstruction of objects through a reasonable collaborative observation model and local feature alignment between collaborative agents, while saving a lot of time in the reconstruction process.

## III. Cooperative Object Reconstruction

Given the set of simultaneously captured images $\{IM\}_{k=1}^{K}$ from $K$ collaborative agents (i.e., cameras), our goal is to recover the 3D model $P_{obj}$ of the object through these images from different viewpoints. To this end, we propose a two-stage pipeline as shown in Fig. 1, including *single-viewpoint object reconstruction* (§III-B) and *a multi-viewpoint concatenation* (§III-C). First, the single-viewpoint object reconstruction method takes each agent to reconstruct a local 3D object (partial 3D information of the target object) from 2D images. Next, the multi-view concatenation method estimates the relative poses by aligning the intrinsic features of collaborative agents in the reconstruction space and thereby integrates all local information into a complete 3D model.

### A. Preliminaries and Core Assumptions

Our method builds on the classic stereo depth estimation [10] to ensure scale consistency across agents. In addition, the binocular depth estimation can also help to realize fast scene reconstruction under fixed-point shooting(i.e., single-viewpoint). It imitates the human eye to estimate the depth. After establishing the correspondence between the points in the left and right views, the 3D coordinate information of the scene can be calculated according to parallax and geometric imaging principles. In this paper, we focus on extracting high-quality local information about the target object from 3D scene representation and then effectively correlating individual agents under large-scale spatial transformations and organizing all local information.

**Assumption.** We assume that the cooperative scenarios of agents are unknown in advance. Thus, the cooperation between agents relies on visual observation, i.e., finding the salience feature of each agent to cooperatively reconstruct a 3D object. Unlike methods [24], [42], we do not enforce the existence of multiple objects with explicit semantics in the environment.

### B. Reconstruction in Single-Viewpoint

Accurate and smooth 3D representation is an important prerequisite for effectively correlating different viewpoints, and also directly determines the reconstruction quality of the target object. To this end, we utilize an advanced stereo-matching network to better guide binocular reconstruction.

Let $IM_k := \{iml_k, imr_k\}$ be the original image pair produced under the $k$-th camera, we hope to recover the scene 3D coordinates $\mathrm{S}_k \in \mathbb{R}^{N\times 3}$ (i.e., scene point cloud) through them. The first is to remove the radial and tangential distortions, and then we perform the stereo rectification for the de-distorted images by the Bouguet algorithm [15]. Note that the camera parameters required in the above process can be obtained through official inquiry or manual calibration [44]. The rectified left and right images $\{\bar{iml}_k, \bar{imr}_k\}$ are row-aligned, that is, the same object point in the real world is located in the same row in both images. After that, we need to find the corresponding point in the right image $\bar{imr}_k$ for the pixel in the left image $\bar{iml}_k$ and calculate the disparity. In this work, instead of using traditional stereo-matching methods, we obtain higher-quality disparity maps end-to-end based on CF-Net [29]. Taking the rectified stereo images $\{\bar{iml}_k, \bar{imr}_k\}$ as input, the network outputs the disparity maps $\{dpl_k, dpr_k\}$ with the same resolution. Finally, We choose one of the left and right views as the datum to calculate the 3D coordinates. Take the left view as an example, for any pixel point $(u, v)$ of $\bar{iml}_k$, the 3D coordinates $[X\ Y\ Z]^T$ of the corresponding object point can be calculated by the following formulas.

$$X = \frac{-B(u-c_x)}{d-(c_x-c_{x'})}, Y = \frac{-B(v-c_y)}{d-(c_x-c_{x'})}, Z = \frac{-Bf}{d-(c_x-c_{x'})} \quad (1)$$

where $d$ denotes the value of the corresponding position of the disparity map $dpl_k$. $c_x, c'_x$ denote the abscissa of the optical center of the left and the right camera in the respective image planes, respectively. $c_y$ is the ordinate of the left camera's optical center in its image plane. The baseline length of the binocular camera is represented by $B$, and $f$ is the focal length.

To further obtain the smooth local point cloud of the target object, we sequentially perform target segmentation and filtering. We first use YOLACT [1] to extract the coarse target local point cloud $\mathrm{E}_k \in \mathbb{R}^{N_e\times 3}$ from the scene point cloud $S_k$. Specifically, since the scene point cloud $S_k$ has a one-to-one correspondence with the pixels in the left view $\bar{iml}_k$, we also use the left view $\bar{iml}_k$ as the input of the network. The network will perform pixel-level object detection on the input image $\bar{iml}_k$ and generate a class label and corresponding Mask for each detected object, so we can

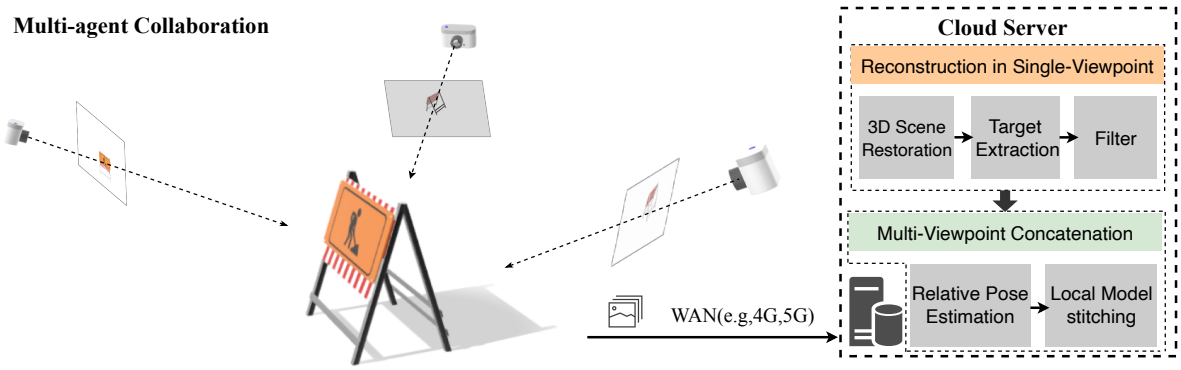


Fig. 1. Several agents detect the target at the same time and upload the observation images from their respective perspectives to the cloud server to complete the object reconstruction. We first recover the 3D scene under each viewpoint and extract the local model belonging to the target object from it. Then, We achieve the concatenation of all local 3D models using the pose estimation parameters generated by aligning the intrinsic features of collaborative agents in the reconstruction space. Note that the figure shows a case where the number of cooperating agents is three.

use this information to easily retrieve and extract the point cloud of the specified object from the scene point cloud. Considering the possible noise in the mask and the noise of the reconstruction process itself, We finally use a statistical filter to smooth the $\mathrm{E}_k$ to obtain the filtered local point cloud of the target object $P_k \in \mathbb{R}^{N_o \times 3}$. The threshold value $h$ of the filter is defined by the Eq. 2. The filter will sequentially calculate the average distance $O_i$ between the $i$-th 3D point in $\mathrm{E}_k$ and its $w$ (custom) nearest neighbors. If $O_i$ is greater than the threshold, the 3D point will be marked as an outlier and removed, otherwise is reserved.

$$h = \bar{O} + m\sqrt{\frac{1}{N_e - 1}\sum_{i=1}^{N_e}(O_i - \bar{O})^2}, \bar{O} = \frac{1}{N_e}\sum_{i=1}^{N_e} O_i \quad (2)$$

where $\bar{O}$ is the average of $O_i$ and $m$ is a custom multiplier.

At this stage, each stereo-pair image can be processed independently to obtain the target local point cloud $P_k$ under the corresponding viewpoint.

### C. Multi-Viewpoint Concatenation

In this section, we describe how we estimate the relative poses of the agents under the task of cooperative object reconstruction. With the estimated parameters, we can easily associate all participating agents and integrate all local 3D point clouds to restore the complete target object. Our key idea to approach the problem is to exploit the mutual visibility between collaborative agents, because we can make sure that some of them are visible to each other with controllable observation positions. Specifically, we introduce a *marker* that is pre-added to each participating camera (or agent) so that it can appear in other camera's view along with the target object as shown in Fig. 2. The advantages of using markers are mainly reflected in the following two aspects: (1) The marker is carefully handcrafted, can be designed to be easier to extract features (with strong texture), and be insensitive to light; (2) Compared with the target object, the overlapping area between the observed markers from two adjacent viewpoints is much larger. Therefore, it is easier to match the makers instead of parts of the target object.

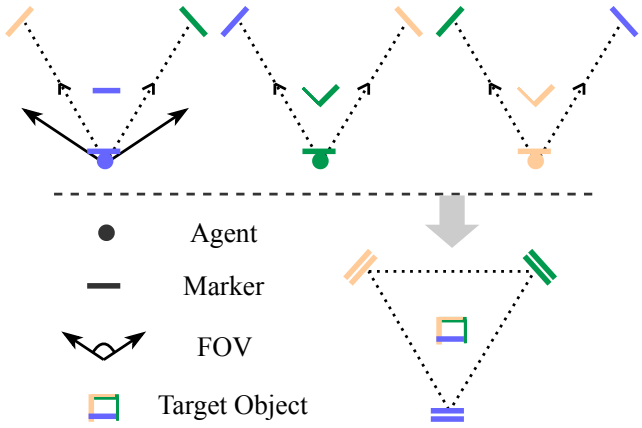


Fig. 2. **Bird's-eye view:** The local 3D models under various viewpoints can be effectively concatenated through the synchronous alignment of markers.

Thereafter, we perform point cloud matching on the marker part to establish a more precise and stable connection between different viewpoints. After the point cloud matching of markers is completed, We build an optimization model to find the best transformation parameters to synchronously align the corresponding markers in the reconstruction space as much as possible. Simultaneously, the parameters can also be used to concatenate the local point clouds of the target object since they represent relative pose transformations between the collaborative agents.

To conduct marker-based point cloud matching, We first crop the complete markers from the original images according to the masks generated by YOLACT [1]. After obtaining the markers' images under all viewpoints, the same marker is grouped and each type of marker represents a class of YOLACT [1]. Then, the A-SIFT [21] is used to extract the key points (features) from markers' images and match these key points. When the coordinates and matching relationship of the key points on the markers' images are determined, we further determine their corresponding coordinates on the original images according to the cropping positions of the markers. Then, the point clouds corresponding to the key

points (i.e. feature point cloud) can be extracted in the scene point clouds. We group them by the classes of the corresponding markers. Finally, a set of matched point clouds is constructed from these feature point cloud groups.

Let $\mathcal{M} := \left\{ \mathrm{M}_k \in \mathbb{R}^{2\times N_m \times 3}, 1 \le k \le K \right\}$ be the set of the matched point clouds, where $\mathrm{M}_k := \left\{ \mathrm{V} \in \mathbb{R}^{N_m \times 3} \subseteq \mathrm{S}_i', \mathrm{Q} \in \mathbb{R}^{N_m \times 3} \subseteq \mathrm{S}_j' \mid i \neq j \right\}$ denotes the feature point cloud groups. We aim to estimate the relative poses of the agents by aligning all feature point cloud groups synchronously, including rotation matrices $\mathrm{R}_k \in SO(3) = \left\{ \mathrm{R} \in \mathbb{R}^{3\times 3} \mid \mathrm{R}\mathrm{R}^T = \mathrm{I}, det(\mathrm{R}) = 1 \right\}$ and translation vectors $\mathbf{t}_k \in \mathbb{R}^3$. Therefore, we first define $\Gamma(\mathrm{M}_k) := \frac{\sum_n^{N_m} \|\mathrm{R}_i \mathrm{V} + \mathbf{t}_i - \mathrm{R}_j \mathrm{Q} - \mathbf{t}_j\|^2}{N_m}$ to denote the Euclidean distance between the $k$-th group of feature point clouds. To achieve synchronous alignment of $K$ groups of feature point clouds, we transform this problem to minimize the sum of their Euclidean distances. Furthermore, we set constraints to avoid possible local optima. Specifically, the case of local optima is that there may be large differences in the degree of alignment between different groups, but the overall alignment is still the best. For this problem, our constraint is that three groups of feature point clouds are arbitrarily selected, and the sum of the Euclidean distances of the two groups must be greater than the third group. The optimization problem can be expressed as follows:

$$\begin{aligned} \min_{\mathrm{R}_i, \mathbf{t}_i} \quad & \sum_{k=1}^{K} \Gamma(\mathrm{M_k}) \\ \text{s.t.} \quad & \Gamma(\mathrm{M_i}) + \Gamma(\mathrm{M_j}) > \Gamma(\mathrm{M_k}), \\ & \forall i, j, k \in \{1, 2, ..., K\}, i \neq j \neq k \end{aligned} \tag{3}$$

In the end, we use Particle swarm optimization (PSO) [19] to solve the above optimization problem. The dimension of the particle is defined as $K * 6$. Under the appropriate population size and particle search space, the algorithm can find the sub-optimal solution in a relatively short time. The obtained parameters are used to perform a coordinate transformation on local point clouds $\{P\}_{k=1}^{K}$ to implement the concatenation.

## IV. Experiments

Our experiments aim to verify the effectiveness of the proposed pipeline. We compare the performance of this framework with the classical 3D reconstruction pipeline (SfM+MVS) and deep learning-based reconstruction methods. The indoor and outdoor scenarios are both considered for a more comprehensive evaluation. In addition, We quantitatively analyze the main factors affecting the performance of the pipeline. Finally, we provide ablation studies and give assessments and suggestions for each critical component of the pipeline.

### A. Experimental setup

**Experiment configuration.** We place three cameras evenly as a circle and the target object is placed at the center point. Thus, the cameras are equidistant from each other and the target. The camera is ZED2 developed by STEREOLABS, with a field of view (FOV) of 110*70, and the range of capture depth can reach as far as 20m. ZED2 can generate 4K (4416*1242 pixel) binocular high-definition images. Moreover, our experiments are conducted on a Ubuntu server with an NVIDIA TESLA V100 GPU (16GB).

**Baselines.** We compare the performance of COLMAP [28] and deep learning-based methods including 3D-R2N2 [5], AttSets [41], Pix2Vox [37], Pixel2Mesh++ [34], and Pix2Vox++ [38] with that of our pipeline. COLMAP is a general-purpose SfM and MVS pipeline with a graphical and command-line interface and is widely used in many real-world applications as a representative work of traditional methods. The selected deep learning-based methods are the state-of-the-art methods in the 3D object reconstruction task, with excellent performance on the well-known public dataset ShapeNet [3].

**Evaluation Configuration.** Our evaluations consist of two phases: *Reconstruction effect* and *Ablation study*. In Reconstruction effect experiments, six real-world objects are recorded from indoor and outdoor scenarios. Note that some baseline solutions are not designed for few-views 3D reconstruction, we obtain more than three views for these methods. It is very hard to have a quantitative comparison for real-world tests; following previous works [2], [23], our comparison is done visually, by observing the structure and texture of generated 3D models.

### B. Results

**Indoor objects.** Fig. 3 shows that we select a cabinet, chair, and kettle as the target objects for reconstruction. All the target objects can be well constructed as 3D cloud points with only 3 views. In addition, we compare our solution with the current mainstream multi-view 3D reconstruction methods. The Colmap is designed for a single camera that collects sequence views from the target objects in 360 degrees. Hence its reconstruction quality increases with the increasing number of input views. Our solution has better reconstruction quality than Colmap even under unequal contrast conditions, i.e., 3 views V.S. 96 views. Most deep learning-based methods focus on constructing the structure of the object, and their outputs are represented as voxels or grids. All these deep learning-based solutions can not obtain promising from images that are directly obtained from real-world scenarios.

**Outdoor objects.** To further evaluate the proposed method under different conditions, we conduct the experiments in an outdoor environment. Test objects include a trashcan, a motorcycle, and a warning cone. Fig. 4 shows that our solution can still obtain a high-quality 3D reconstruction and is not affected by the strong light in the outdoor environment. The performance of Colmap has not fluctuated much in the outdoor environment, and the reconstruction results are close to the indoor environment. However, deep learning-based methods are sensitive to light. When the target object has relatively complex structures (e.g., motorcycle), these models completely fail in the outdoor environment.

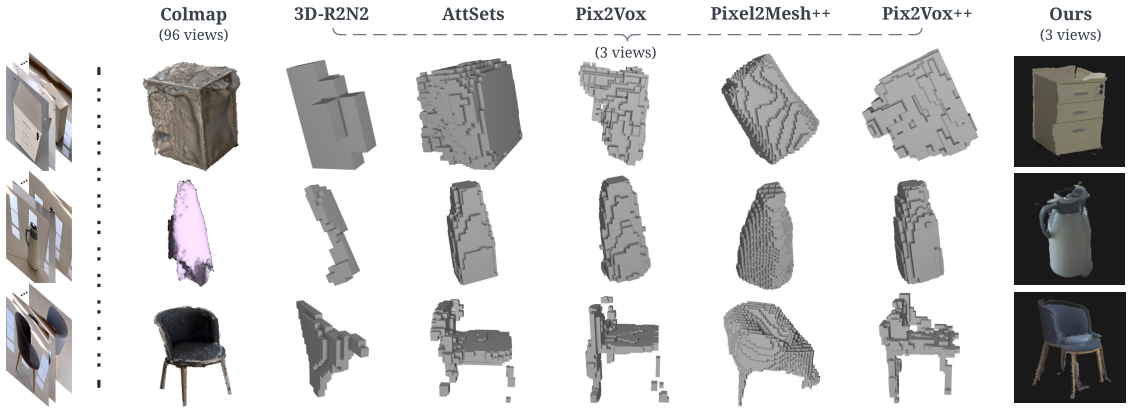


Fig. 3. Comparison of reconstruction effects on indoor objects

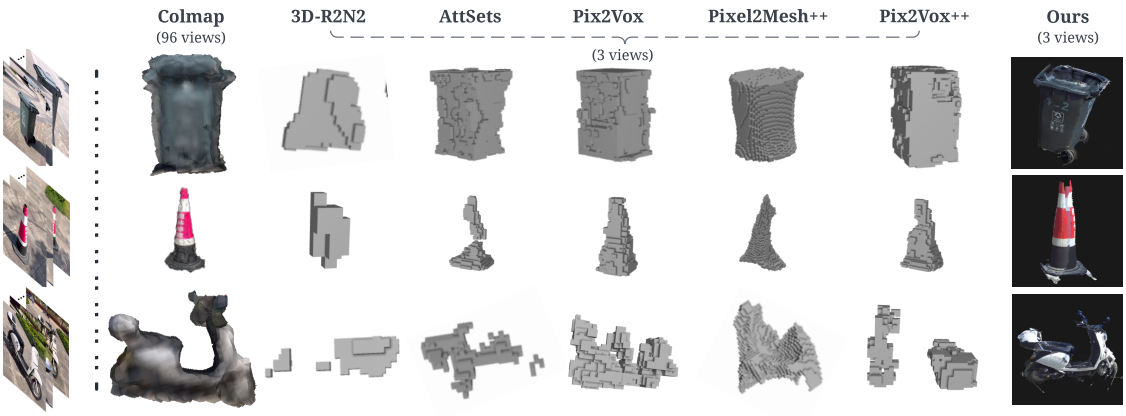


Fig. 4. Comparison of reconstruction effects on outdoor objects

TABLE I

RECONSTRUCTION QUALITY AND TIME-CONSUMING COMPARISON OF DIFFERENT METHODS.

| | **COLMAP** (96 views) | **Pix2Vox++** etc. (3 views) | **Ours** (3 views) |
|---|---|---|---|
| Average time(s) | 7956 | 17 | 50 |
| Structure | ✓ | ✓ | ✓ |
| Texture | ✓ | N/A | ✓ |
| Scale | ✓ | N/A | ✓ |

**Effectiveness.** We also evaluate the effectiveness of our proposed solution in terms of *execution time* and *supported visual features*. Table I shows that deep learning-based solutions outperform the other two solutions in terms of time-consuming. However, these methods are not able to reconstruct the texture and scale information of the target objects. Our proposed method achieves the best reconstruction quality, but the execution latency is about 1 minute. In future work, we develop a faster optimization algorithm to replace PSO to achieve online 3D reconstruction.

### C. Robustness Analysis

We define the average distance between point cloud pairs of well-matched marker regions $Bias := \frac{\sum_{k=1}^{K} \Gamma(\mathrm{M}_k)}{K}$ to reflect the reconstruction effect of the proposed pipeline quantitatively. Smaller $Bias$ means better alignment of the markers and thus smaller connection gaps between the local point clouds.

**The Impact of Viewpoint Allocations.** The relative pose relationship between cooperating agents is an important factor affecting its robustness. The ideal position is that the cooperative cameras are distributed in a regular polygon so that the pose changes between every two adjacent viewpoints are the same. To evaluate the robustness of our solution under non-ideal observation positions, as shown in Table II, we change the cameras' placement and compare the performance of reconstructing a carton under different viewpoint assignments (including ideal and non-ideal positions). The experimental results demonstrate that our proposed solution is almost unaffected by the variety of poses between cooperating agents. Our proposed pipeline works even if the position and pose distributions among the cooperating agents are not so canonical. As a result, our proposed solution is usable as long as the markers can be observed by the corresponding agents.

**Specifications of Markers.** The size of the markers determines the upper limit of the size of objects that our method can reconstruct. First, the YOLCAT algorithm requires the marker to occupy a sufficient proportion of pixels in the image to successfully detect it. In addition, we also need to ensure that the ASIFT algorithm can extract feature points in

TABLE II

NON-UNIFORM VIEWPOINT ASSIGNMENT TESTS: $[T_x, T_y, T_z]$ AND $[\alpha, \beta, \gamma]$ DENOTE THE TRANSLATION AND ROTATION OF THE CAMERA RELATIVE TO THE REFERENCE CAMERA ($C_1$), RESPECTIVELY. THE $G_1$ (I.E. IDEAL DISTRIBUTION) SERVED AS THE REFERENCE GROUP FOR THE EXPERIMENT.

| | | Cam-Pan $[T_x, T_y, T_z]$ | Cam-Rotation $[\alpha, \beta, \gamma]$ | Bias | Time(s) |
|---|---|---|---|---|---|
| $C_1$ | | $[0, 0, 0]$ | $[0, 0, 0]$ | - | - |
| $\mathbf{G_1}$ | $C_2$ | $[1.2, 2.08, 0]$ | $[0, 0, 120]$ | 57.6 | 49 |
| | $C_3$ | $[-1.2, 2.08, 0]$ | $[0, 0, 240]$ | | |
| $\mathbf{G_2}$ | $C_2$ | $[1.2, 2.08, 0]$ | $[0, 0, 120]$ | 56.9 | 50 |
| | $C_3$ | $[-1.39, 1.39, 0]$ | $[0, 0, 270]$ | | |
| $\mathbf{G_3}$ | $C_2$ | $[1.1, 2, 0]$ | $[0, 0, 120]$ | 57.2 | 53 |
| | $C_3$ | $[-1.3, 2.2, 0]$ | $[0, 0, 240]$ | | |
| $\mathbf{G_4}$ | $C_2$ | $[1.2, 2.08, 0]$ | $[0, 0, 120]$ | 58.1 | 47 |
| | $C_3$ | $[-2.08, 2.08, 0]$ | $[0, 0, 250.5]$ | | |
| $\mathbf{G_5}$ | $C_2$ | $[1.2, 2.08, 0.1]$ | $[20, 10, 120]$ | 57.4 | 51 |
| | $C_3$ | $[-2.08, 2.08, 0.2]$ | $[30, -10, 250.5]$ | | |

TABLE III

COMPARISON OF THE MAXIMUM SIZE OF OBJECTS THAT CAN BE RECONSTRUCTED BY MARKERS OF DIFFERENT SIZES.

| **Marker size** (mm×mm) | **Detection distance**(m) accuracy ≥50% | **Feature points** (pcs) | **Maximum radius** (m) |
|---|---|---|---|
| 148×209 | 0∼4.72 | ≥19 | 1.36 |
| 274×463 | 0∼6.86 | ≥21 | 1.98 |
| 413×694 | 0∼9.84 | ≥25 | 2.84 |
| 507×891 | 0∼13.21 | ≥29 | 3.81 |

the marker area. Therefore, we select markers with different sizes (take white paper as an example), count the critical values separately, and infer the maximum object size that the proposed pipeline can allow. The resolution of the image is fixed at 2208*1242, and the results are shown in Table III. The detection distance refers to the straight-line distance between the camera and the markers of the collaborative agents. To this end, the choice of markers should be based on the actual situation.

### *D. Ablation studies*

**Parallax optimization.** Obtaining the disparity map of the target object is an essential step for the entire pipeline. We, therefore, study the impact of various parallax generation algorithms on the pipeline. Fig. 5 shows the comparison of the reconstruction process using CF-Net and SGM respectively. SGM [11] is the most widely used stereo-matching algorithm in commercial software. The experimental results demonstrate that the disparity map obtained by SGM has relatively poor quality. Although the hollow part can be culled during the filtering process, resulting in the destruction of the integrity of the object reconstruction, the smoothness of the obtained marker directly affects the final stitching result.

**Point cloud pairing.** A reasonable pairing of point clouds is a prerequisite for the pipeline to achieve marker alignment. We also tested the method of uniformly selecting or randomly sampling several point clouds from the marker point cloud set for pairing and solving the optimal parameters. Fig. 6 shows a set of comparison results. Compared with the feature-level pairing used in the pipeline, the biases

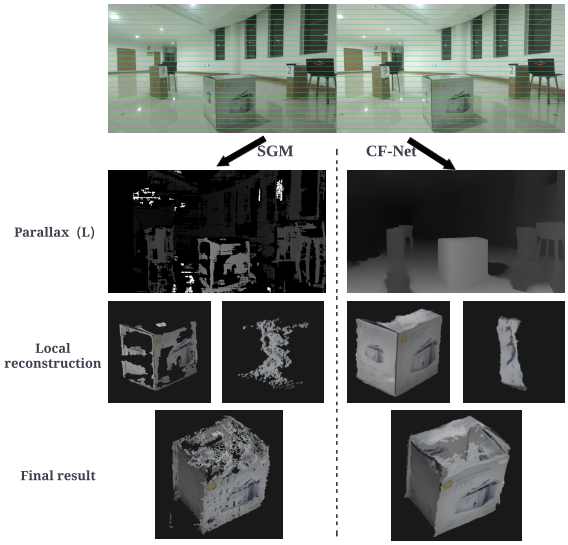


Fig. 5. Comparison of the work of the proposed pipeline under two different parallax acquisition methods.


Fig. 6. Stitching test under different point cloud pairing methods.

of the other two methods are $30.82\%$ and $16.68\%$ higher respectively. However, in practical systems, they can be used as a temporary alternative when the feature-matching algorithm fails and the peer-to-peer pairing fails.

## V. CONCLUSIONS

In this paper, we develop a novel object reconstruction pipeline based on multi-agent collaboration, which can restore the 3D model of the target object in a scene from observed images from a few agents. First, the pipeline will perform local 3D restoration of the target object in each viewpoint, and then estimate the relative poses between all viewpoints and perform point cloud stitching based on the estimated parameters to obtain the final reconstruction result. Experiments show that our proposed pipeline can effectively deal with various objects in the real world, and has significant advantages over other methods in the comprehensive evaluation of reconstruction quality and time-consuming.

A primary limitation that can cause reconstructing failures is when the instance segmentation network cannot accurately identify and extract targets from the image, e.g. unseen objects. In future work, it would be interesting to use limited agents to efficiently reconstruct objects in a complex scenario. To this end, the proposed algorithm needs to plan a strategy to obtain the images with the smallest efforts for reconstructing multiple scattered objects.

## REFERENCES


[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019.

[2] Mingwei Cao, Liping Zheng, Wei Jia, and Xiaoping Liu. Joint 3d reconstruction and object tracking for traffic video analysis under iov environment. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3577–3591, 2020.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Yu Chen, Shuhan Shen, Yisong Chen, and Guoping Wang. Graph-based parallel large scale structure from motion. *Pattern Recognition*, 107:107537, 2020.

[5] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[6] Viktoriia Vitalievna Evdokimova, Maksim Vitalyevich Petrov, Marina Alexandrovna Klyueva, Eugene Yu Zybin, Vladislav Viktorovich Kos' yanchuk, Irina Borisovna Mishchenko, VM Novikov, Nikolay Ivanovich Sel'vesyuk, Egor Ivanovich Ershov, Nikolay Alexandrovich Ivliev, et al. Deep learning-based video stream reconstruction in mass-production diffractive optical systems. *Computer Optics*, 45(1):130–141, 2021.

[7] Amnon Geifman, Yoni Kasten, Meirav Galun, and Ronen Basri. Averaging essential and fundamental matrices in collinear camera settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6021–6030, 2020.

[8] Xiaodong Gu, Weihao Yuan, Zuozhuo Dai, Chengzhou Tang, Siyu Zhu, and Ping Tan. Dro: Deep recurrent optimizer for structure-from-motion. *arXiv preprint arXiv:2103.13201*, 2021.

[9] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1578–1604, 2019.

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[11] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[12] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, and Rajiv Suman. Substantial capabilities of robotics in enhancing industry 4.0 implementation. *Cognitive Robotics*, 1:58–75, 2021.

[13] San Jiang, Cheng Jiang, and Wanshou Jiang. Efficient structure from motion for large-scale uav images: A review and a comparison of sfm tools. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:230–251, 2020.

[14] Yiwei Jin, Diqiong Jiang, and Ming Cai. 3d reconstruction using deep learning: a survey. *Communications in Information and Systems*, 20(4):389–413, 2020.

[15] Annika Kuhl. Comparison of stereo matching algorithms for mobile robots. *Centre for Intelligent Information Processing System*, pages 4–24, 2005.

[16] Alex Locher, Michal Perdoch, and Luc Van Gool. Progressive prioritized multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3244–3252, 2016.

[17] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. *arXiv preprint arXiv:2206.05737*, 2022.

[18] Rachel Macrorie, Simon Marvin, and Aidan While. Robotics and automation in the city: a research agenda. *Urban Geography*, 42(2):197–217, 2021.

[19] Federico Marini and Beata Walczak. Particle swarm optimization (pso). a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149:153–165, 2015.

[20] Mateusz Michalkiewicz, Sarah Parisot, Stavros Tsogkas, Mahsa Baktashmotlagh, Anders Eriksson, and Eugene Belilovsky. Few-shot single-view 3-d object reconstruction with compositional priors. In *European Conference on Computer Vision*, pages 614–630. Springer, 2020.

[21] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM journal on imaging sciences*, 2(2):438–469, 2009.

[22] Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2018.

[23] Charalambos Poullis and Suya You. 3d reconstruction of urban areas. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 33–40. IEEE, 2011.

[24] Shengyi Qian, Linyi Jin, and David F Fouhey. Associative3d: Volumetric reconstruction from sparse views. In *European Conference on Computer Vision*, pages 140–157. Springer, 2020.

[25] Redmond R Shamshiri, Cornelia Weltzien, Ibrahim A Hameed, Ian J Yule, Tony E Grift, Siva K Balasundram, Lenka Pitonakova, Desa Ahmad, and Girish Chowdhary. Research and development in agricultural robotics: A perspective of digital farming. 2018.

[26] Ammar Abdul Ameer Rasheed, Mohammed Najm Abdullah, and Ahmed Sabah Al-Araji. A review of multi-agent mobile robot systems applications. *International Journal of Electrical & Computer Engineering (2088-8708)*, 12(4), 2022.

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[28] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[29] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021.

[30] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[31] Leslie Ching Ow Tiong, Dick Sigmund, and Andrew Beng Jin Teoh. 3d-c2ft: Coarse-to-fine transformer for multi-view 3d reconstruction. *arXiv preprint arXiv:2205.14575*, 2022.

[32] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[33] Xiang Wang, Chen Wang, Bing Liu, Xiaoqing Zhou, Liang Zhang, Jin Zheng, and Xiao Bai. Multi-view stereo in the deep learning era: A comprehensive revfiew. *Displays*, 70:102102, 2021.

[34] Chao Wen, Yinda Zhang, Chenjie Cao, Zhuwen Li, Xiangyang Xue, and Yanwei Fu. Pixel2mesh++: 3d mesh generation and refinement from multi-view images. *arXiv preprint arXiv:2204.09866*, 2022.

[35] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1042–1051, 2019.

[36] Xin Wen, Junsheng Zhou, Yu-Shen Liu, Hua Su, Zhen Dong, and Zhizhong Han. 3d shape reconstruction from 2d images with disentangled attribute flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3803–3813, 2022.

[37] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2690–2698, 2019.

[38] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020.

[39] Zhen Xing, Yijiang Chen, Zhixin Ling, Xiangdong Zhou, and Yu Xiang. Few-shot single-view 3d reconstruction with memory prior contrastive network. *arXiv preprint arXiv:2208.00183*, 2022.

[40] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019.

[41] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *International Journal of Computer Vision*, 128(1):53–73, 2020.

[42] Jingyu Yang, Ji Xu, Kun Li, Yu-Kun Lai, Huanjing Yue, Jianzhi Lu, Hao Wu, and Yebin Liu. Learning to reconstruct and understand indoor scenes from sparse views. *IEEE Transactions on Image Processing*, 29:5753–5766, 2020.

[43] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018.

[44] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[45] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26:305–364, 2017.