2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Robust Map Fusion with Visual Attention Utilizing Multi-agent Rendezvous

Jaein Kim[1,3], Dong-Sig Han[2,3], and Byoung-Tak Zhang[1,2,3]

***Abstract*— The map fusion for multi-robot simultaneous localization and mapping (SLAM) consistently combines robot maps built independently into the global map. An established approach to map fusion is utilizing rendezvous, which refers to an encounter between multiple agents, to calculate the transformation into the global map. However, previous works using rendezvous have a limitation in that they are unreliable for certain circumstances, where the amount of agent observations or overlapping landmarks is limited. This work proposes a novel map fusion system which robustly fuses local maps in challenging rendezvous that lack shared information. Our system utilizes the single visual perception from rendezvous and estimates the relative pose between agents with the DOPE. Then our scheme transforms local maps with an estimated relative pose and predicts the misalignment from approximated maps by utilizing the attention mechanism of the vision transformer. Comparisons with the Hough transform-based method show that ours is significantly better when the overlap between local maps is insufficient. We also verify the robustness of our system against a similar real-world scenario.**

I. Introduction

The simultaneous localization and mapping (SLAM), where a mobile robot localizes itself and builds a consistent map simultaneously [1], becomes more challenging in the multi-robot domain. In such a domain, an algorithm should consider the uncertainties caused by distributed information [2]. A map fusion, one of the solutions for multi-robot SLAM, fuses independently built local maps into a global map. Thus it should compute the optimal transformations from local frames into the global frame [3].

The relative pose between multi-robot agents is one of the crucial elements of map fusion [4]. Some fusion methods compute map transformations only utilizing visual features extracted from given local maps [5]–[7]. However, these methods are very limited, because they require sufficient overlap between local maps which cannot be assured [3].

A map fusion can utilize a relative pose between agents when *rendezvous* occurs, in which agents encounter each other. This setup makes map fusion more robust and is used in diverse multi-robot SLAM methods [8]–[13]. However, most previous works have limitations in that they require multiple observations to compute reliable alignments toward the global frame, or still utilize a feature matching for revision. The ideal map fusion, however, should acquire the global map from a single encounter between agents regardless of the amount of shared information, as depicted in Figure 1.

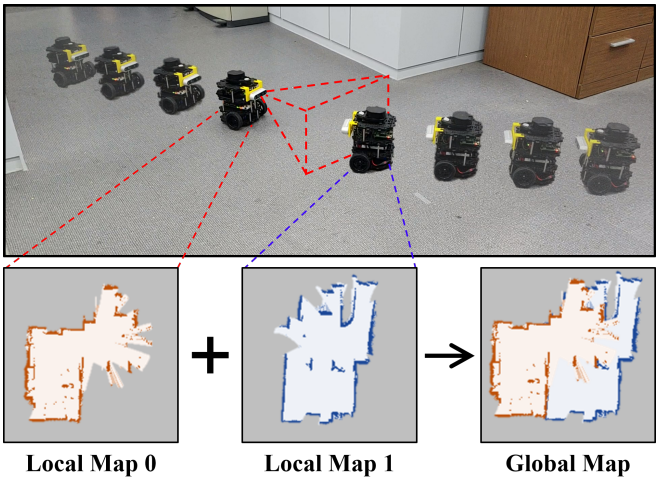

Fig. 1. The rendezvous situation of multiple agents. A map fusion system using rendezvous merges local maps into a global map by utilizing the relative pose measurement. Ideally, local maps can be transformed into the global frame even with a single observation and minimal shared map information by perfect relative pose measurement.

This work suggests a novel map fusion system, which aligns local maps initially with a single observation and refines any possible misalignment with the method appropriate for the rendezvous situation. We implement important functionalities of the system with deep neural networks to ensure reliable performance when available information such as agent encounters or map overlap is restricted.

Our system estimates the relative pose of an observed agent from an RGB image and computes the approximate alignment of local maps from the estimated pose. To refine the error caused by pose measurement noise, it also estimates the misalignment between local maps by utilizing perturbed local features from each input map. Notably, it does not use classic feature matching as in previous works, because only minimal overlaps may exist in the rendezvous situation. Instead, the attention mechanism, which models a relation between ordered features [14], is proposed to estimate the alignment error.

This work contains the following contributions:

- We apply the DOPE [15] to our pose measurement model. It makes our method acquire an approximate

*This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)/25%, 2021-0-02068-AIHub/25%, 2022-0-00951-LBA/25%, 2022-0-00953-PICA/25%)]

[1]Interdisciplinary Program in Neuroscience, Seoul National University
[2]Dept. of Computer Science and Engineering, Seoul National University
[3]Artificial Intelligence Institute, Seoul National University
{jykim, dshan, btzhang}@bi.snu.ac.kr

relative pose with only a single observation.

- We claim that the attention mechanism is suitable to estimate the misalignment between local maps if the identical landmarks are deficient. Our revision model estimates the perturbation by embedding related features, including non-shared ones, with the vision transformer [16].
- We train the above modules on large-scale datasets we made from scratch, and justify the necessity of each module. Finally, through the real-world multi-robot experiment, we verify that our map fusion system outperforms the existing methods in extreme conditions where shared features between local maps are restricted.

II. Related Works

A. Map Fusion Utilizing Rendezvous

A map fusion method using rendezvous finds an optimal spatial transformation toward the global map by utilizing relative information between agents at the encounter. How it fuses local maps with noisy observations and optimizes the global map varies for each research.

For example, the method in [8] uses Extended Kalman Filter for each agent and optimizes fused maps by matching identical landmarks. A robot trajectory is formulated as a pose graph and constrained with relative measurements from each encounter in [10] and [11]. Especially in [11], loop closures in fused maps are detected by the RANSAC feature matching. The work in [12] tackles relative localization from rendezvous and trajectory tracking simultaneously. The map fusion system in [13] stacks non-static features to identify identical features from local maps. A common limitation in the above previous works is that multiple observations or sufficient identical features between maps are required to get a reliable fused map.

B. Attention Mechanism & Transformer

The attention mechanism refers to the computational perception which focuses on a task-specific subspace of input information, inspired by human perception [17]. According to [14], the attention model, which represents the relation between sequences of input and output, is proposed for the sequential domain first in [18]. It uses a recurrent encoder-decoder architecture to map embedded sequences into relations between input and output sequences.

However, an attention mechanism based on a recurrent framework requires sequential attention computation, which causes computational inefficiency [14]. The transformer [19] parallelizes multiple self-attention calculations by the *multi-head attention*, and has shown outstanding performance in terms of long-range dependency, parallel processing, and scalability.

The vision transformer (ViT) applies the multi-head attention to the image space by dividing it into disjoint patches and adding a whole image representation token. It has earned renown in recent computer vision research and has already been applied to various vision tasks [20]–[23].

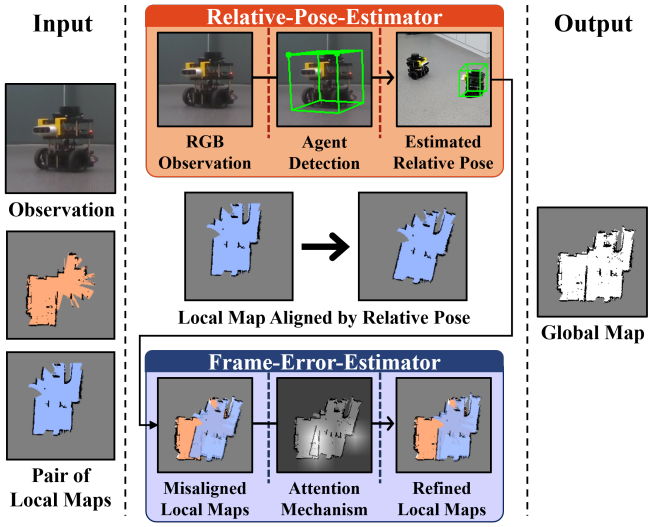

Fig. 2. The abstract scheme of our system. The RELATIVE-POSE-ESTIMATOR detects an agent projected on input observation and estimates its relative pose. Then the estimated pose is used to align the local map from the observed agent. The FRAME-ERROR-ESTIMATOR encodes attention between visual features from the aligned local map pair and estimates any misalignment in it. We visualize this attention as the self-attention map at the bottom of this figure using the Attention Rollout [24]. Our system reverses the estimated misalignment of the map from the observed agent and fuses local maps into the global map using Algorithm 1.

III. Methods

This section explains the entire flow of our map fusion system. We highlight that there are two key modules in our system, which are depicted in Figure 2.

A. Problem Notations and Setups

Notations in this paper are as follows:

- R_i: An agent in multi-robot system. $i \in \{0, 1\}$. R_0 captures R_1 without loss of generality.
- I: A single RGB image capturing R_1.
- $M_i \in \mathbb{R}^{H \times W}$: An occupancy grid map built by R_i.
- $T \in SE(2)$: The true transformation from M_1 to M_0.
- $\theta_i \in \mathbb{R}^3$: A pose of R_i in M_i.
- $\theta_{\text{relative}} \in \mathbb{R}^3$: A relative pose from R_0 to R_1.
- $\theta_{\text{error}} \in \mathbb{R}^3$: A misalignment between local maps aligned by only using θ_{relative}.

This study suggests the map fusion system which fuses maps from two agents in the multi-robot system when the agents encounter. We assume that the agents explore independently in the same single-floor indoor environment with unknown initial positions and build their occupancy grid maps consistently using arbitrary SLAM algorithms. The system operates when one of the agents captures another on its RGB camera sensor and succeeds in recognizing it.

The pose θ_i of each agent on its map and the relative pose θ_{relative} are represented by three parameters: translations on

the XY plane and rotation on the Z axis:

$$\begin{bmatrix}\Delta_x & \Delta_y & \Delta_\psi\end{bmatrix}^T \Longleftrightarrow \begin{bmatrix}\cos\Delta_\psi & -\sin\Delta_\psi & \Delta_x \\ \sin\Delta_\psi & \cos\Delta_\psi & \Delta_y \\ 0 & 0 & 1\end{bmatrix}, \tag{1}$$

where $\Delta_x, \Delta_y \in \mathbb{R}$, and $\Delta_\psi \in (-\pi, \pi]$.
As Equation 1, the pose with bounded orientation corresponds to the coordinate value of the matrix in $SE(2)$ group. Thus we designate pose values and corresponding $SE(2)$ matrices by the same symbol without distinction.

B. System Overview

The proposed map fusion system is initialized when the RELATIVE-POSE-ESTIMATOR detects R_1 from I and estimates θ_{relative}. The system computes $\tilde{T}$, the approximate of T by using θ_{relative}, as Equation 2:

$$\tilde{T} = \theta_0 \ \theta_{\text{relative}} \ \theta_1^{-1}. \tag{2}$$

Due to the noise in relative pose estimation, the frame of $\tilde{T}M_1$ is aligned to the one of M_0 with local perturbation, although they should be ideally identical frames. The FRAME-ERROR-ESTIMATOR estimates this frame perturbation θ_{error}, and $\hat{T}$, the refined estimation of T, is computed as Equation 3:

$$\hat{T} = \theta_{\text{error}}^{-1}\tilde{T}. \tag{3}$$

Finally, the map fusion heuristic merges refined local maps, M_0 and $\hat{T}M_1$, while prioritizing conflicting pixel information between maps. This function is in Algorithm 1.

C. Robot Pose Estimation on Single Scenery Image

The RELATIVE-POSE-ESTIMATOR module is based on the DOPE, which encodes features from an agent capturing the image using convolution layers. Encoded features are mapped to predict the belief map B and vector field V projected on input image space. B represents belief states of vertices and a centroid composing the bounding cuboid of the agent. V represents directions from vertices toward the centroid. These convolution layers are trained using $\mathcal{L}_{\text{pose}}$ in

Algorithm 1 Local Maps Fusion with Conflicts Prioritization

Input: Local Maps M_0, $\hat{T}M_1$
Output: Fused Global Map M

$M \leftarrow (m_{ij}) \in \mathbb{R}^{H\times W}$
for all $m_{ij} \in M$ **do**
  if $(M_0)_{ij}$ or $(\hat{T}M_1)_{ij}$ are "OCCUPIED" **then**
    $m_{ij} \leftarrow$ "OCCUPIED"
  else if $(M_0)_{ij}$ and $(\hat{T}M_1)_{ij}$ are "UNKNOWN" **then**
    $m_{ij} \leftarrow$ "UNKNOWN"
  else
    $m_{ij} \leftarrow$ "EMPTY"
  end if
end for
return M

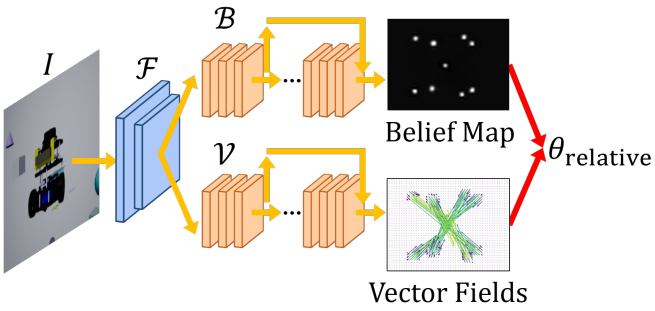

Fig. 3. The flow from I to θ_{relative} via RELATIVE-POSE-ESTIMATOR. Orange processes are flows of convolution networks, and red processes are deterministic heuristics and algorithms to output θ_{relative}. Belief maps and vector fields are visualized by being stacked into single images.

Equation 4:

$$\begin{aligned}\mathcal{L}_{\text{pose}}(I, B, V) =& \|\mathcal{B}(\mathcal{F}(I)) - B\|_2^2 \\ &+ \|\mathcal{V}(\mathcal{F}(I)) - V\|_2^2,\end{aligned} \tag{4}$$

where $\mathcal{F}(\cdot)$ outputs shared features, $\mathcal{B}(\cdot)$ outputs predicted belief map in $\mathbb{R}^{9\times H\times W}$, and $\mathcal{V}(\cdot)$ outputs predicted vector filed in $\mathbb{R}^{16\times H\times W}$.

The module decides the bounding cuboid from the predicted B and V, and the pose from camera frame to agent is computed using the Perspective-n-Points algorithm [25]. The model outputs a pose in 3D space, but only translations on the XY plane and rotations on the Z axis are transferred to the following process as θ_{relative}. The entire flow of RELATIVE-POSE-ESTIMATOR is depicted in Figure 3.

We used the data synthesis tool called NViSII [26] to obtain a sufficient amount of photorealistic images to train the model. It renders RGB scenes of a robot with random poses. To mitigate reality gaps, random backgrounds and obstacles are rendered together for augmentation.

The RELATIVE-POSE-ESTIMATOR only requires a single RGB image as input for localization of the observed agent. Thus it makes our pose estimation method more robust than previous works, but the noise from the estimated pose remains to be handled. With only a single rendezvous, we cannot formulate the map fusion task as a reliable optimization problem or exploit the classic feature matching algorithm due to the restriction of shared landmarks between local maps. However, the RELATIVE-POSE-ESTIMATOR still provides $\tilde{T}$, which is an initial point at the near of T for our method to search θ_{error}.

D. Coordinate Frame Error Estimation Using Attention

The FRAME-ERROR-ESTIMATOR module assumes that the perturbation from T to $\tilde{T}$ is in a predictable region, and utilizes the pair of $(M_0, \tilde{T}M_1)$ as an input to estimate θ_{error}. The ViT architecture receives the input map pair as a two-channel image and regresses θ_{error} with a simple MLP. This process is depicted in Figure 4.

During its training process, the pair of misaligned M_0 and $\tilde{T}M_1$ should be given at every step. Thus we uniformly sampled θ_{error} from the error distribution of relative pose

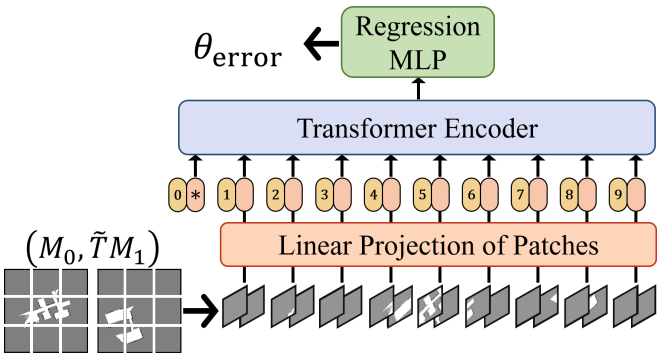

Fig. 4. The architecture of FRAME-ERROR-ESTIMATOR. An ordered pair of local maps, $(M_0, \tilde{T}M_1)$, is divided into disjoint two-channel patches with identical spatial coordinates. These patches are projected into linear representations sequentially and encoded with an attention mechanism. The transformer encoder sums the projected representations and the encoded features into a single feature vector. Finally, the regression MLP receives the feature vector and outputs θ_{error}. This illustration is based on [16].

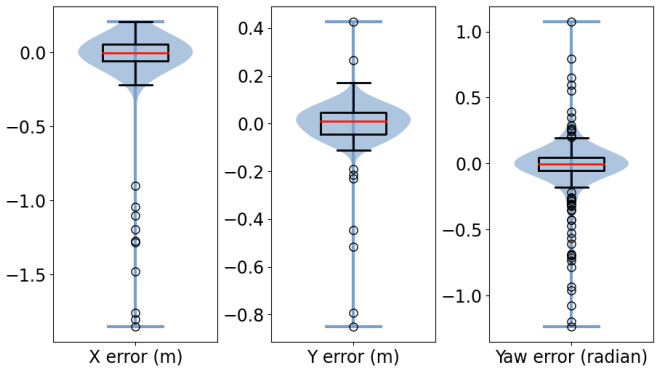

Fig. 5. Box plots and violin plots of the generalization error of the RELATIVE-POSE-ESTIMATOR for each element of pose.

estimation and perturbed the ground truth M_1 with θ_{error}. The range of sampling distribution was set to cover the extreme generalization errors from the RELATIVE-POSE-ESTIMATOR. See Section IV-A for details.

The noteworthy detail for FRAME-ERROR-ESTIMATOR is the choice of a loss function. Because the elements of the regression target have different scales and ranges, the model struggles to learn the coordinate value of the $SE(2)$ group directly. To naturally balance such $SE(2)$ target θ_{error}, the module $\mathcal{M}$ learns corresponding $se(2)$ algebra instead as Equation 5 [27]:

$$\begin{aligned}&\mathcal{L}_{\text{error}}(M_0, \tilde{T}M_1, \theta_{\text{error}}) \\ &\quad = \frac{1}{2} d(\xi, \theta_{\text{error}})^T \Sigma^{-1} d(\xi, \theta_{\text{error}}),\end{aligned} \tag{5}$$

where $\xi = \mathcal{M}(M_0, \tilde{T}M_1)$, and $d(\xi, \theta) = \log\left(\exp(\xi)\theta^{-1}\right)$. Here we omit the conversion between a matrix and a vector in the group or algebra, and use an identity matrix as Σ for simplicity. We also add the unsupervised learning term to regularize with heuristics. Because $\tilde{T}M_1$ is transformed from M_0 as θ_{error}, $\theta_{\text{error}}M_0$ should be aligned with $\tilde{T}M_1$ ideally. Thus the regularization term is designed as Equation 6:

$$\mathcal{L}_{\text{reg}}(M_0, \tilde{T}M_1) = \|\mathcal{M}(\theta M_0, \tilde{T}M_1)\|_2, \tag{6}$$

where $\theta = \exp\left(\mathcal{M}(M_0, \tilde{T}M_1)\right)$.
This spatial transformation is differentiable using the method from [28]. The total loss function $\mathcal{L}_{\text{frame}}$ for FRAME-ERROR-ESTIMATOR is as Equation 7:

$$\mathcal{L}_{\text{frame}} = \mathcal{L}_{\text{error}} + \omega\mathcal{L}_{\text{reg}}, \quad \omega \in [0, 1]. \tag{7}$$

The dataset of local map pairs has to be made from scratch to imitate realistic occupancy grids in rendezvous situations. We used indoor floorplans from the HouseExpo dataset [29] to build sufficient Gazebo simulation environments. Meshes of walls for given indoor floorplans were generated using a script modified from [30]. Multiple agents explored the generated environments while executing SLAM, and local maps were recorded when a rendezvous occurred.

Overall, the FRAME-ERROR-ESTIMATOR estimates the coordinate frame error between local maps, even when matching map features are minimal. We claim this is possible because the proposed method uses the attention mechanism, which can weigh relations between distinct visual features from misaligned local maps. Due to this property, the FRAME-ERROR-ESTIMATOR is suitable for rendezvous situations.

IV. EXPERIMENTS

In this section, we test the suggested modules to verify their robustness to the challenging rendezvous, both in the test dataset and the real-world setup. During the data generation and the real-world experiment, we used the TurtleBot3 Burger platform with Intel Realsense Depth Camera D435 as an agent.

A. Relative Pose Estimation Performance

The RELATIVE-POSE-ESTIMATOR was tested on the set of 1000 photorealistic synthetic images capturing an agent. We plot the distribution of estimated pose error in Figure 5. The plot shows that the error distributions are zero-centered and their variances are small enough to support that the relative pose estimation is very accurate for most cases. However, extrema also exist, which still provides evidence of the true pose but harms the quality of estimation largely. Based on these results, we can say that the FRAME-ERROR-ESTIMATOR needs to robustly cover the extreme cases of pose error.

B. Comparison and Ablation Studies of Frame Error Estimation on Local Map Pairs Dataset

We tested the FRAME-ERROR-ESTIMATOR on the test set from our dataset, collected independently with the training set using the same generation procedure. For the baseline method, we implemented the algorithm from [5], which we denote as `Hough`. It transforms local maps into spectral signals using Hough transformation and exhaustively searches the $SE(2)$ parameter that maximizes the correlation between map spectral signals.

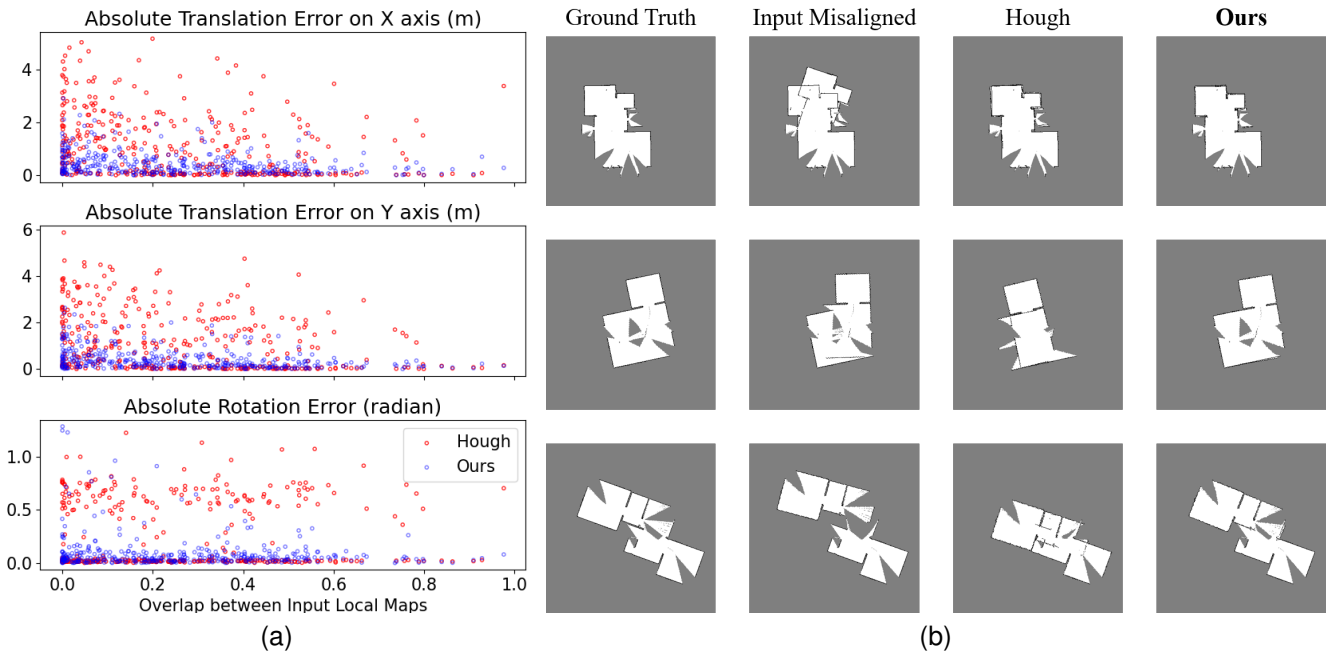

Fig. 6. (a) The scatter plots for the absolute estimation errors of the FRAME-ERROR-ESTIMATOR and the baseline `Hough` method against the local maps' overlap ratio. (b) The map fusion results of the proposed and the baseline on the test set. From the left, each column corresponds to the ground truth, misaligned input map pair, fusion result of the Hough, and the result of our suggested method.

We also conducted ablation studies by varying the image encoding architecture or loss terms of FRAME-ERROR-ESTIMATOR. In Table I, the `ViT` denotes whether a model encodes map pairs using a vision transformer or convolution network. The `Reg` denotes whether the loss term in Equation 6 is used, and the `Lie` denotes whether the regression target is Lie algebra or naive pose value.

We recorded the absolute error of θ_{error} from every variation and baseline in Table I. It shows that the models regressing Lie algebra with a transformer encoder have the best performance. When comparing with CNN variations, models encoding input with ViT and regressing Lie algebra target can fully utilize the scalability of a transformer to learn the inductive bias of merging local maps and outperform others including the baseline.

The detailed comparison with the baseline is in Figure 6. The plot in Figure 6 (a) shows that the translation error of the baseline method tends to increase as the overlap between input local maps decrease. This tendency is shown clearly in Figure 6 (b). The `Hough` fuses local maps accurately when overlaps between input local maps are distinct. However, it overestimates occupied regions when local maps lack overlaps, and the translation error exceedingly increases. It shows the known issue of existing map fusion methods that match shared visual features. On the other hand, our method shows robust frame error estimation performance regardless of the amount of shared features between local maps.

C. System Performance in Multi-robot Scenarios

Finally, we tested the performance of the entire map fusion system in real-world multi-robot scenarios. We implemented this system using the Robot Operating System (ROS), and the system operated in the centralized workstation with GPU. Agents were manually controlled to navigate toward the rendezvous region while building their local maps with GMapping [31]. The system received local maps and an observation image when agents encountered each other. This scenario and inputs from it are shown in Figure 7.

TABLE I
MEAN AND 95% CONFIDENCE INTERVAL OF THE ABSOLUTE ERROR OF ESTIMATED θ_{error} ON TEST SET

ViT	Reg	Lie	$\lvert\Delta_x\rvert$ (m)	$\lvert\Delta_y\rvert$ (m)	$\lvert\Delta_\psi\rvert$ (radian)
✗	✗	✗	0.586 ± 0.050	0.591 ± 0.054	0.130 ± 0.021
✗	✗	✓	0.562 ± 0.047	0.534 ± 0.050	0.167 ± 0.021
✗	✓	✗	0.700 ± 0.054	0.692 ± 0.054	0.149 ± 0.019
✗	✓	✓	0.556 ± 0.051	0.570 ± 0.051	0.175 ± 0.020
✓	✗	✗	0.999 ± 0.060	0.964 ± 0.059	0.394 ± 0.023
✓	✗	✓	$\mathbf{0.425 \pm 0.047}$	0.441 ± 0.050	$\mathbf{0.097 \pm 0.018}$
✓	✓	✗	0.638 ± 0.057	0.632 ± 0.054	0.138 ± 0.024
✓	✓	✓	0.438 ± 0.046	$\mathbf{0.426 \pm 0.048}$	$\mathbf{0.097 \pm 0.018}$
Hough			1.096 ± 0.125	1.175 ± 0.131	0.261 ± 0.033

In this experiment, we compared our system with two existing methods as baselines. The first method is the same as we used in Section IV-B, which we denote as the `Hough`. The second method from [32] extracts ORB features from local maps and estimates the transformation parameters with RANSAC. We denote this method as the `ORB`. Miscellaneous configurations on the experiment are in Table II.

The results from baselines and our method are in Figure 7. Although the RELATIVE-POSE-ESTIMATOR estimated a quite

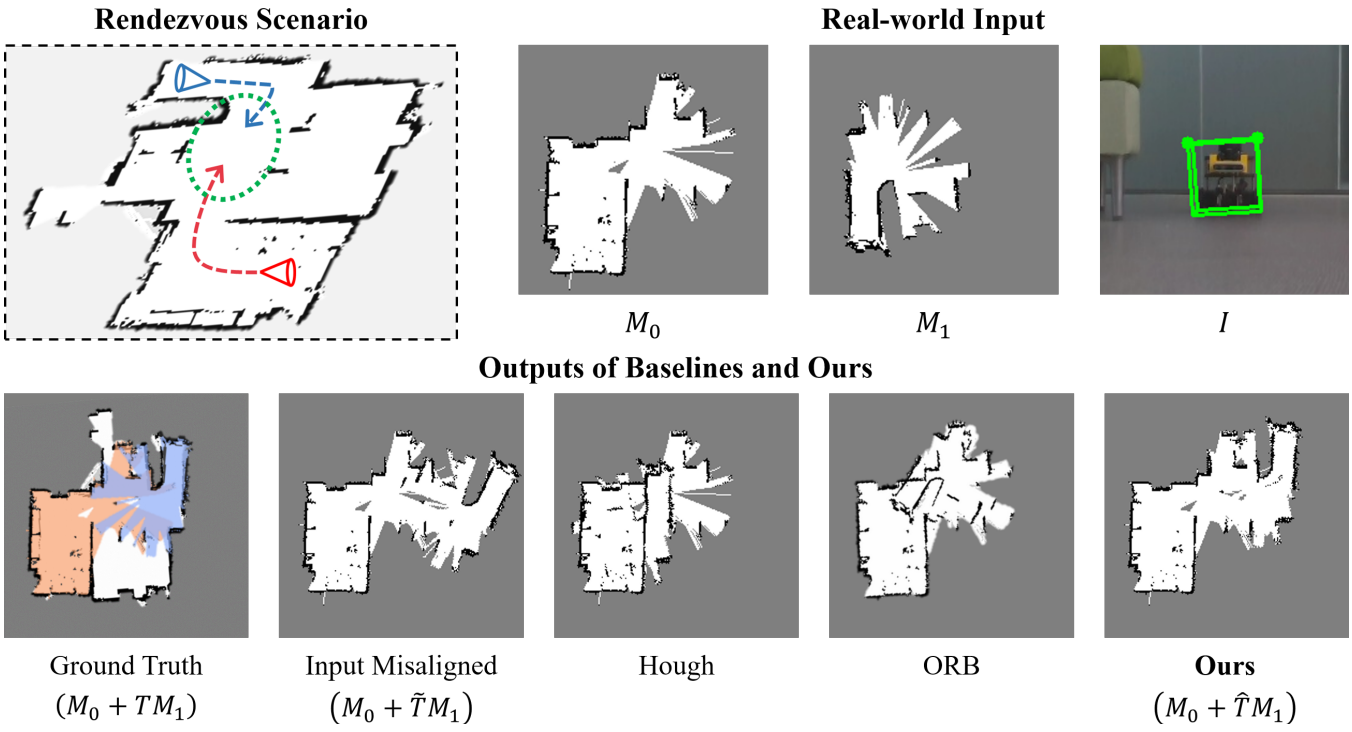

Fig. 7. The real-world map fusion scenario and its results on baselines and ours. The red cone is the initial pose of R_0, and the red arrow is its trajectory. The blue ones are R_1's and vice versa. When agents encounter each other in the rendezvous area, which is denoted as a green circle, the system receives agents' local maps and the observation image. We overlay the estimated bounding cuboid on the input observation I. On the bottom, the global maps of ground truth, using only relative pose, other baseline methods, and our FRAME-ERROR-ESTIMATOR are shown. To visualize the true fused map and how much overlap between local maps, we paint and overlay M_0 (red) and M_1 (blue) on the map of the entire environment at the ground truth item.

TABLE II
REAL-WORLD SCENARIO CONFIGURATIONS

Map Configurations			Ground Truth T		
Size	Resolution	Overlap	Δ_x	Δ_y	Δ_ψ
384×384	0.05 m/pixel	0.174	6.0 m	-5.8 m	180°

accurate bounding cuboid on the observed agent, the fused maps only using pose information $(M_0 + \tilde{T}M_1)$ were still inaccurate. Baseline methods, as expected, failed to fuse local maps, because given local maps have minimal overlap between themselves. With the FRAME-ERROR-ESTIMATOR, however, the proposed map fusion system robustly refined the misalignment. Though the global map from ours $(M_0 + \hat{T}M_1)$ had a minor error due to the inconsistency between local maps independently built in the real world, the result shows that our method solely exploited clues unlike other baselines, and robustly solved the map fusion task in the challenging rendezvous situation.

V. CONCLUSION & DISCUSSION

This paper proposed the map fusion system utilizing rendezvous that mitigates the issues of existing map fusion methods, where a sufficient amount of shared information, such as the number of observations or matching landmarks between local maps, is required. We implemented the RELATIVE-POSE-ESTIMATOR by utilizing the DOPE and made relative pose estimation possible with only a single image. To refine the approximate transformation computed with a relative pose, we implemented the FRAME-ERROR-ESTIMATOR, which encodes transformed local maps utilizing the attention mechanism with ViT. This scheme makes the model embed features from local maps with minimal overlaps. We verified the robustness of the suggested system through experiments on diverse levels, from the synthesized dataset to the real-world scenario.

In future works, the proposed system can be extended further to broader setups or environments. For instance, applying our system to 3D maps, which are used widely in real-world applications, is considerable. It can be done by extending attention models for 3D data points, which have been studied recently by applying a transformer network [33], [34]. Currently, our pose estimation module only assumes an indoor environment with mild conditions on agent observation like lighting or occlusion. It may increase uncertainties of relative pose estimation in the real world, especially in the outdoor environment. Quantifying these uncertainties in pose estimation for robotics [35] can help future works maintain robustness in diverse setups.

ACKNOWLEDGMENTS

We thank Heebin Yoo, Inwoo Hwang, Min Whoo Lee, Chung-Yeon Lee and Beom-Jin Lee for helpful comments and discussions.

REFERENCES

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[2] L. E. Parker, "Distributed intelligence: Overview of the field and its application in multi-robot systems." in *AAAI fall symposium: regarding the intelligence in distributed intelligent systems*, 2007, pp. 1–6.

[3] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-robot simultaneous localization and mapping: A review," *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.

[4] I. Andersone, "Heterogeneous map merging: State of the art," *Robotics*, vol. 8, no. 3, p. 74, 2019.

[5] S. Carpin, "Fast and accurate map merging for multi-robot systems," *Autonomous robots*, vol. 25, no. 3, pp. 305–316, 2008.

[6] J.-L. Blanco, J. González-Jiménez, and J.-A. Fernández-Madrigal, "A robust, multi-hypothesis approach to matching occupancy grid maps," *Robotica*, vol. 31, no. 5, pp. 687–701, 2013.

[7] H. Lee and S. Lee, "Grid map merging with insufficient overlapping areas for efficient multi-robot systems with unknown initial correspondences," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 1427–1432.

[8] X. S. Zhou and S. I. Roumeliotis, "Multi-robot slam with unknown initial correspondence: The robot rendezvous case," in *2006 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2006, pp. 1785–1792.

[9] L. Carlone, M. K. Ng, J. Du, B. Bona, and M. Indri, "Rao-blackwellized particle filters multi robot slam with unknown initial correspondences and limited communication," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 243–249.

[10] B. Kim, M. Kaess, L. Fletcher, J. Leonard, A. Bachrach, N. Roy, and S. Teller, "Multiple relative pose graphs for robust cooperative mapping," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 3185–3192.

[11] J. Dong, E. Nelson, V. Indelman, N. Michael, and F. Dellaert, "Distributed real-time cooperative localization and mapping using an uncertainty-aware expectation maximization approach," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 5807–5814.

[12] M. W. Mehrez, G. K. Mann, and R. G. Gosine, "An optimization based approach for relative localization and relative tracking control in multi-robot systems," *Journal of Intelligent & Robotic Systems*, vol. 85, no. 2, pp. 385–408, 2017.

[13] Y. Jang, C. Oh, Y. Lee, and H. J. Kim, "Multirobot collaborative monocular slam utilizing rendezvous," *IEEE Transactions on Robotics*, 2021.

[14] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.

[15] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning*. PMLR, 2018, pp. 306–316.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[21] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 502–10 511.

[22] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[23] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5791–5800.

[24] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 4190–4197.

[25] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[26] N. Morrical, J. Tremblay, Y. Lin, S. Tyree, S. Birchfield, V. Pascucci, and I. Wald, "Nvisii: A scriptable tool for photorealistic image generation," 2021.

[27] V. Peretroukhin and J. Kelly, "Dpc-net: Deep pose correction for visual localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2424–2431, 2017.

[28] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[29] T. Li, D. Ho, C. Li, D. Zhu, C. Wang, and M. Q.-H. Meng, "Houseexpo: A large-scale 2d indoor layout dataset for learning-based algorithms on mobile robots," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5839–5846.

[30] S. Curtis, "Ros package for creating gazebo environments from 2d maps." [Online]. Available: https://github.com/shilohc/map2gazebo

[31] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.

[32] J. Hörner, "Map-merging for multi-robot system," Bachelor's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Prague, 2016.

[33] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.

[34] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268.

[35] G. Shi, Y. Zhu, J. Tremblay, S. Birchfield, F. Ramos, A. Anandkumar, and Y. Zhu, "Fast uncertainty quantification for deep object pose estimation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5200–5207.