Autonomous Robots (2023) 47:635–648
https://doi.org/10.1007/s10514-023-10086-9

Uncertainty-aware correspondence identification for collaborative perception

Peng Gao[1] · Qingzhao Zhu[2] · Hao Zhang[3]

Received: 25 February 2022 / Accepted: 12 January 2023 / Published online: 6 February 2023

Abstract
Correspondence identification is essential for multi-robot collaborative perception, which aims to identify the same objects in order to ensure consistent references of the objects by a group of robots/agents in their own fields of view. Although recent deep learning methods have shown encouraging performance on correspondence identification, they suffer from two shortcomings, including the inability to address non-covisibility and the inability to quantify and reduce uncertainty to improve correspondence identification. To address both issues, we propose a novel uncertainty-aware deep graph matching method for correspondence identification in collaborative perception. Our new approach formulates correspondence identification as a deep graph matching problem, which identifies correspondences based on deep graph neural network-based features and explicitly quantify uncertainties in the identified correspondences under the Bayesian framework. In addition, we design a novel loss function that explicitly reduces correspondence uncertainty and perceptual non-covisibility during learning. Finally, we design a novel multi-robot sensor fusion method that integrates the multi-robot observations given the identified correspondences to perform collaborative object localization. We evaluate our approach in the robotics applications of collaborative assembly, multi-robot coordination and connected autonomous driving using high-fidelity simulations and physical robots. Experiments have shown that, our approach achieves the state-of-the-art performance of correspondence identification. Furthermore, the identified correspondences of objects can be well integrated into multi-robot collaboration for object localization.

Keywords Correspondence identification · Deep graph matching · Uncertainty quantification · Non-covisibility · Collaborative perception

1 Introduction

Collaborative robotics, including multi-robot systems (Brambilla et al., 2013; Chung et al., 2018; Reily et al., 2020) and human-robot collaboration (Matsumoto & Riek, 2019; Reily et al., 2018), has been widely studied over the past decades due to its effectiveness and flexibility to address large-scale collaborative tasks. Collaborative perception is a fundamental capability in collaborative robotics for robots and other agents including humans in a collaborative team to share information of the surrounding environment thus achieving shared situational awareness among the teammates. Collaborative perception has been widely applied in a variety of real-world applications including human-robot collaborative assembly (Hietanen et al., 2020; Gao et al., 2020b), multi-robot mapping and navigation (Acevedo et al., 2020; Yue et al., 2020), and connected autonomous driving (Guo et al., 2019; Wei et al., 2018). Correspondence identification is defined as a problem to identify the same objects observed by multiple agents in their own fields of view, which is considered an essential component to enable collaborative perception (Frey et al., 2019; Gao et al., 2020a; Tian et al., 2019). For example, as illustrated by Fig. 1, when a collaborative robot assists a human worker who wears an augmented reality (AR) headset to assemble a chair, they need to identify the correspondence of the chair parts in order to ensure

✉ Peng Gao
gaopeng@umd.edu

Qingzhao Zhu
zhuqingzhao@mines.edu

Hao Zhang
hao.zhang@cs.umass.edu

[1] University of Maryland, College Park, USA

[2] Colorado School of Mines, Golden, USA

[3] University of Massachusetts Amherst, Amherst, USA

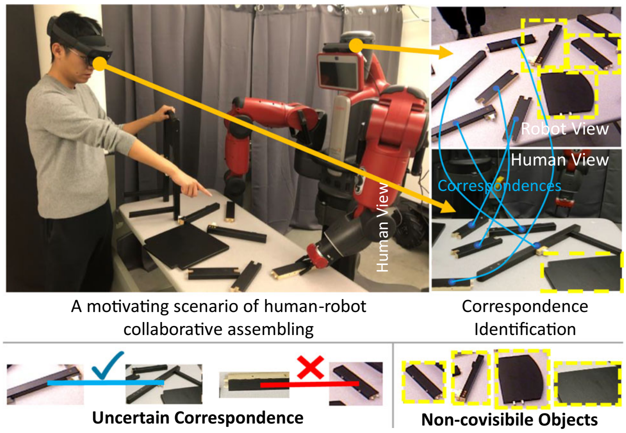


Fig. 1 This example motivates correspondence identification in collaborative perception in the application of human-robot collaborative assembly. When a collaborative robot assists a human worker who wears an augmented reality (AR) headset to assemble a chair, they must identify the correspondence of the chair parts in order to ensure that both the robot and the human correctly refer to the same object used in the assembling operations. We propose a novel Bayesian deep graph matching method for correspondence identification with the capability of explicitly reducing correspondence uncertainty and perceptual non-covisibility in collaborative perception

that both the robot and the human correctly refer to the same object.

Given its importance, many techniques have been developed to address correspondence identification, e.g., based on visual object reidentification (Zhao et al., 2016, 2019b) and learning-free graph matching (Chang et al., 2017; Cho et al., 2010). Recently, deep learning has attracted significant attention for identifying correspondences in collaborative perception due to its ability to learn from data and its robustness to noise. For example, through learning visual features using convolution neural networks (CNN) (Shi et al., 2016; Yu et al., 2018), the methods for object reidentification identified the same objects in different frames and from different perspectives (Jin et al., 2020; Khatun et al., 2020; Quispe & Pedrini, 2019; Wang et al., 2019a; Voigtlaender et al., 2019). By encoding spatial relationships of the objects using graph neural networks (GNN) (Fey et al., 2018; Veličković et al., 2018), deep graph matching was designed to learn graph similarities (Wang et al., 2019b; Zhang & Lee, 2019) and graph representations (Fey et al., 2019; Jiang et al., 2019) for correspondence identification. Compared with the deep feature learning, deep graph matching is able to explicitly integrate both visual and spatial information of the objects for improved identification.

However, the current state-of-the-art deep graph matching methods suffer from two key shortcomings that have not been yet addressed for collaborative perception. First, the previous approaches are not able to quantify and reduce the *uncertainty* in identified correspondences. Uncertainty is always expected in collaborative perception, e.g., due to sensor resolution limit and measurement noise (Gal & Ghahramani, 2015). Without the capability of explicitly quantifying and addressing uncertainties during learning, deep graph matching is not robust to noisy observations (Kendall et al., 2018). The second shortcoming stems from *non-covisibility*, which is defined as the challenge that not all objects are observed by all agents due to occlusion and limited field of view (Fig. 1). Non-covisibility makes objects in the observations that are acquired from different perspectives to have no correspondence, which has not been addressed by current deep graph matching methods.

We propose a novel Bayesian deep graph matching method for correspondence identification, with the capability of explicitly modeling and addressing uncertainty and non-covisibility in collaborative perception. We first represent each observation acquired by an agent as a graph. Nodes of the graph encode visual appearances of the detected objects in the observation and the edges denote spatial relationships among the objects in the robot's field of view. Then, given two graphs built from observations by a pair of agents, we formulate correspondence identification as a problem of Bayesian deep graph matching. Furthermore, we introduce a novel loss function that models and reduces non-covisibility and uncertainty in the unidentified correspondences during learning.

The key contribution of this paper is the introduction of the first Bayesian deep graph matching approach that models and addresses uncertainty and non-covisibility for correspondence identification in multi-agent collaborative perception. Specific novelties include:

– We introduce a novel approach for Bayesian deep graph matching, which integrates graph matching with Bayesian deep learning to solve correspondence identification. Our approach explicitly models and quantifies uncertainty in the identified object correspondences, thus improving the interpretability of deep graph matching.
– We introduce a new loss function that reduces correspondence uncertainty and perceptual non-covisibility, which improves the robustness of correspondence identification to noisy observations during collaborative perception.

A preliminary conference version of this work was published at Robotics Science and System 2021 (Gao & Zhang, 2021). We extend the previous conference work as follows. First, in Sect. 3, we propose a follow-up algorithm on collaborative object localization, which is based on the results obtained from the proposed deep graph matching method. Second, we perform a case study in a new scenario on connected autonomous driving in Sect. 4, in order to evaluate our novel approach on correspondence identification as well as collaborative object localization. Finally, we discuss about the future study direction to improve the current approach on multi-robot collaborative perception in Sect. 5.

2 Related work

2.1 Correspondence identification

Conventional methods for correspondence identification can be grouped into three categories, based on visual appearances for object reidentification, spatial relationships for learning-free graph matching, and pairwise association for multi-view synchronization. The first category of methods calculate the similarity of two observations based on local (Engel et al., 2014), global (Zhao et al., 2016), or semantic features (Zhao et al., 2019b). The second category of methods use the spatial similarity among objects using, e.g., distances between the objects in pairwise graph matching (Cho et al., 2010; Leordeanu & Hebert, 2005), angular relationships of objects in hypergraph matching (Nguyen et al., 2015; Suh et al., 2015), spatial relationships built by four or more objects in clique matching (Nie et al., 2015), and a combination of multiple spatial relationships (Chang et al., 2017). The third category of methods recognize object correspondences by enforcing the circle-consistent constraints in multiple views (Fathian et al., 2020), e.g., based on convex relaxation (Boyd et al., 2011), spectral relaxation (Maset et al., 2017) and graph clustering (Yan et al., 2016).

The conventional methods require that the appearance and spatial pattern of objects must be unique, which are not robust to the perception uncertainty caused by occlusion, noisy data and model bias. Recently, regularized graph matching method is proposed (Gao et al., 2020a), which addresses the observation uncertainty by adding regularization terms into the graph matching formulation. However, this method cannot address the uncertainty in the graph matching model, and is not able to quantify the correspondence uncertainty caused by the perception uncertainty.

2.2 Deep graph matching

Deep graph matching has attracted attention to address correspondence identification in recent years. By aggregating the local visual-spatial information around objects through GNN, deep graph matching learns the similarity between the local visual-spatial embeddings of the objects (Wang et al., 2019b; Zhang & Lee, 2019). The identified correspondence can be improved by designing representative graphs (Jiang et al., 2019) or by removing the correspondences violating neighborhood consensus (Fey et al., 2019). The accuracy of deep graph matching can be improved by incorporating combinatorial solvers (Rolínek et al., 2020), and the efficiency can be improved by decomposing large graphs into small parts (Lou et al., 2020). Deep graph matching outperforms traditional learning-free graph matching methods due to its ability to learn from data and its robustness to noise. Compared with deep reidentification methods, deep graph matching methods encode additional spatial information of the objects, thus improving the representability.

2.3 Uncertainty quantification

Recent deep learning studies have also focused on Bayesian learning frameworks for GNN to quantify the uncertainty in different domains. The type of the uncertainty obtained from Bayesian GNN includes aleatoric uncertainty of the data and epistemic uncertainty of the learning model (Kendall & Gal, 2017), vacuity and dissonance uncertainty from subjective logic perspective (Fey et al., 2019), variance (Gal & Ghahramani, 2016) and entropy (Malinin & Gales, 2018).

The techniques to quantify the uncertainty can mainly be divided into two categories, including non-Bayesian and Bayesian techniques. The most well-known non-Bayesian uncertainty quantification technique is deep ensemble, which makes averaged prediction given a collection of parallel networks (Fort et al., 2019; Lakshminarayanan et al., 2017). The shortcoming of the non-Bayesian methods includes the lack of interpretability and computational expense (running multiple models at the same time). Bayesian-based techniques focus on modeling the distribution of network parameters for uncertainty quantification, including Markov Chain Monte Carlo (MCMC) (Kupinski et al., 2003), Bayes by backprop (BBB) (Blundell et al., 2015) and Monte Carlo Dropout (MC dropout) (Gal & Ghahramani, 2016). The Bayesian-based techniques are widely used in various applications, such as using Bayesian GNN with Dirichlet prior (Malinin & Gales, 2018; Zhang & Lee, 2019) and Gaussian prior (Ryu et al., 2019) for node classification (Zhang et al., 2019a), edge prediction (Zhang et al., 2016) and graph classification (Zhao et al., 2019a).

Given the promising performance of using GNN to represent single observations, there exists no Bayesian learning frameworks for deep graph matching to address correspondence identification in collaborative perception. In addition, previous deep graph matching methods assume that all objects in the source observation are also present in the target observation, which are not applicable to correspondence identification with non-covisible objects. The approach proposed in this paper explicitly addresses the challenges of both uncertainty and non-covisibility in deep graph matching for correspondence identification in collaborative perception.

3 Approach

Notation. Matrices are represented as boldface capital letters, e.g., $\mathbf{M} = \{\mathbf{M}_{i,j}\} \in \mathcal{R}^{n \times m}$, with $\mathbf{M}_{i,j}$ denoting the element in the i-th row and j-th column of $\mathbf{M}$. Vectors are denoted as boldface lowercase letters $\mathbf{v} \in \mathcal{R}^n$ and scalars are denoted as lowercase letters.

3.1 Problem formulation

We propose to formulate correspondence identification in collaborative perception as a deep graph matching problem. Given an observation that's acquired by a robot, we represent it as an undirected graph $\mathcal{G}(\mathbf{V}, \mathbf{A}, \mathbf{E})$. The node matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]^\top \in \mathcal{R}^{n \times d_v}$ denotes the central locations of the objects detected in the observation, where $\mathbf{v}_i \in \mathcal{R}^{d_v}$ is the location of the i-th object and n is the number of objects. The attribute matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]^\top \in \mathcal{R}^{n \times d_a}$ encodes appearance features of these objects, where $\mathbf{a}_i \in \mathcal{R}^{d_a}$ denotes the feature vector of the i-th object. The edge matrix $\mathbf{E} = \{\mathbf{E}_{i,j}\} \in \mathcal{R}^{n \times n}$ denotes the pairwise adjacency of the nodes. The edges are generated by Delaunay triangulation given the coordinates of objects. Thus, if $\mathbf{v}_i$ and $\mathbf{v}_j$ are connected, $\mathbf{E}_{i,j} = ||\mathbf{v}_i - \mathbf{v}_j||_2$ is computed as the distance between $\mathbf{v}_i$ and $\mathbf{v}_j$, otherwise it is zero.

Given this graph representation, we compute the local embeddings of the objects, which capture the neighborhood visual-spatial information around the objects. The local embeddings are computed by $\mathbf{H} = \Psi(\mathbf{A}, \mathbf{E}) = \{\mathbf{h}_i\}^n$, where Ψ is a GNN that is defined as follows:

$$\mathbf{h}_i^l = \sigma(\mathbf{W}^l \mathbf{h}_i^{l-1} + \sum_{j \in \mathcal{N}(i)} \Phi^l(\mathbf{E}_{i,j}) \cdot \mathbf{h}_j^{l-1}) \tag{1}$$

where $\mathbf{W}$ denotes the trainable parameter of GNN, $\mathcal{N}(i)$ denotes the neighborhood objects of the i-th object, $\Phi(\mathbf{E}_{i,j})$ denotes the trainable B-spline kernel function, which uses graph edges connected to the i-th robot to compute the weight of its neighborhood objects for local information aggregation, σ denotes the non-linear function ReLu, and $l \in \{1, 2, \ldots, L\}$ is the number of layers in the forward process of the GNN. $\mathbf{h}_i^l$ denotes the features of objects in different layers. Since we set $L = 2$, $\mathbf{h}_i^2 \in \mathcal{R}^{32}$, $\mathbf{h}_i^1 \in \mathcal{R}^{256}$ and the initial embedding is defined as $\mathbf{h}_i^0 = \mathbf{a}_i$. Thus, $\mathbf{H} = \{\mathbf{h}_i^2\}^n$

In collaborative perception, observations acquired by a pair of robots are represented as two graphs $\mathcal{G}(\mathbf{V}, \mathbf{A}, \mathbf{E})$ and $\mathcal{G}'(\mathbf{V}', \mathbf{A}', \mathbf{E}')$, respectively. We calculate their respective embedding vectors $\mathbf{H}$ and $\mathbf{H}'$ using Eq. (1). Then, the visual-spatial similarity of $\mathcal{G}$ and $\mathcal{G}'$ can be computed as follows:

$$\mathbf{S} = \mathbf{H}\mathbf{H}'^\top = \Psi(\mathbf{A}, \mathbf{E})\Psi^\top(\mathbf{A}', \mathbf{E}') \tag{2}$$

where $\mathbf{S} = \{\mathbf{S}_{i,i'}\}^{n \times n'}$ denotes the similarity matrix with $\mathbf{S}_{i,i'}$ indicating the similarity between the i-th object in graph $\mathcal{G}$ and the i'-th object in $\mathcal{G}'$. Since local embeddings may not be sufficiently distinct when objects have similar local visual-spatial structures, we improve the similarity matrix $\mathbf{S}$ as follows:

$$\mathbf{S} = \mathbf{H}\mathbf{H}'^\top + \varphi(\mathbf{D}) \tag{3}$$

where φ denotes a multi-layer perceptron that is computed as the concatenation of two linear functions with a ReLu non-linear function, and $\mathbf{D}$ denotes the measurement of neighborhood consensus (Fey et al., 2019), which is computed by $\mathbf{D}_{i,j} = \mathbf{Z}_{i,:} - \mathbf{Z}'_{j,:}$ with $\mathbf{Z} = \Psi(\mathbf{A}, \mathbf{E})$ and $\mathbf{Z}' = \Psi(\mathbf{S}^\top\mathbf{A}, \mathbf{S}^\top\mathbf{E}\mathbf{S})$ based on Eq. (1). The intuition is as follows. If the similarity based on local embeddings (Eq. 2) between two graphs $\mathcal{G}$ and $\mathcal{G}'$ can result in correct correspondences (e.g., a large similarity indicates a correct correspondence), when the visual-spatial information of $\mathcal{G}'$ is replaced with the information of $\mathcal{G}$ given the correspondence (e.g,. replacing $\mathbf{A}'$ by $\mathbf{S}^\top\mathbf{A}$), the embedding of $\mathcal{G}$ and the new embedding of $\mathcal{G}'$ should be the same. Otherwise, the difference $\mathbf{D}$, as a measurement of the neighborhood consensus, between the two embeddings of $\mathcal{G}$ and $\mathcal{G}'$ is used to update the similarity matrix.

Then, correspondence identification is formulated as a graph matching problem as follows:

$$\arg\max_{\mathbf{Y}} \mathbf{S}^\top\mathbf{Y} \quad \text{s.t. } \mathbf{Y}\mathbf{1}_{n' \times 1} \leq \mathbf{1}_{n \times 1}, \mathbf{Y}^\top\mathbf{1}_{n \times 1} \leq \mathbf{1}_{n' \times 1} \tag{4}$$

where $\mathbf{Y} = \{\mathbf{Y}_{ii'}\}$ denotes the correspondence matrix, with $\mathbf{Y}_{ii'} = 1$ meaning that the i-th object in $\mathcal{G}$ corresponds to the i'-th object in $\mathcal{G}'$, and $\mathbf{1}$ is a vector with all ones. Equation (4) aims to maximize the overall similarity of objects' embedding given the correspondence matrix $\mathbf{Y}$. The constraints are used to guarantee one-to-one correspondences by enforcing each row and column in $\mathbf{Y}$ to at most have one element equal to 1. Gradient-decent methods can be used to solve Eq. (4), e.g., using the Sinkhorn algorithm (Zhang et al., 2019b; Fey et al., 2019) that is efficient and strict with one-to-one correspondence constraint.

3.2 Quantifying uncertainty in correspondence identification

Uncertainty always exists in robot perception. We propose a Bayesian deep graph matching method that re-designs deep graph matching under the Bayesian learning framework to quantify uncertainty in correspondence identification.

We represent the trainable parameter $\mathbf{W}$ in a distribution form instead of taking fixed values. Given a set of N training instances $\mathcal{X} = \{\mathcal{G}_i^*, \mathcal{G}_{i'}^{*'}\}^N$ with ground truth $\mathcal{Y} = \{\mathbf{Y}_i^*\}^N$, $\mathbf{W}$ is computed as:

$$p(\mathbf{W}|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \mathbf{W})p(\mathbf{W})}{p(\mathcal{Y}|\mathcal{X})} \tag{5}$$

where $p(\mathbf{W}|\mathcal{X}, \mathcal{Y})$ is the posterior distribution of $\mathbf{W}$ estimated from its prior distribution $p(\mathbf{W})$. Given $p(\mathbf{W}|\mathcal{X}, \mathcal{Y})$, the inference process is defined as follows:

$$p(\mathbf{Y}|\mathcal{G},\mathcal{G}',\mathcal{X},\mathcal{Y}) = \int_{\mathbf{W}\in\Omega} p(\mathbf{Y}|\mathcal{G},\mathcal{G}',\mathbf{W})p(\mathbf{W}|\mathcal{X},\mathcal{Y})d\mathbf{W} \tag{6}$$

Under our framework of Bayesian learning, $p(\mathbf{Y}|\mathcal{G},\mathcal{G}',\mathcal{X},\mathcal{Y})$ represents the correspondence matrix $\mathbf{Y}$ in a distribution form, rather than taking fixed values through marginalizing over the posterior $p(\mathbf{W}|\mathcal{X},\mathcal{Y})$. $p(\mathbf{Y}|\mathcal{G},\mathcal{G}',\mathbf{W})$ denotes the probability of $\mathbf{Y}$ given the pair of graphs $\mathcal{G},\mathcal{G}'$ as input and the model parameter $\mathbf{W}$.

Directly computing the integral in Eq. (6) requires to integrate over all the parameter space Ω, which is intractable for the gradient descent-based inference. In order to address this challenge, we adopt the dropout variance inference (Gal & Ghahramani, 2016) to obtain the approximated posterior distribution $q(\mathbf{W})$ instead of $p(\mathbf{W}|\mathcal{X},\mathcal{Y})$ by minimizing the Kullback-Leibler divergence:

$$\min_{\theta} KL(q_{\theta}(\mathbf{W})||p(\mathbf{W}|\mathcal{X},\mathcal{Y})) = \min_{\theta} \int_{\mathbf{W}\in\Omega} q_{\theta}(\mathbf{W}) \log \frac{q_{\theta}(\mathbf{W})}{p(\mathbf{W}|\mathcal{X},\mathcal{Y})} \tag{7}$$

where $\theta = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_N\}$ denotes the variational parameter with $\mathbf{M}_i$ denoting the deep graph matching network's parameters without dropout operations, and N denotes the number of layers in the network.

During training, we sample $\mathbf{W}_i$ from $q_{\theta}(\mathbf{W})$ using dropout as follows:

$$\begin{aligned}\mathbf{W}_i &= \mathbf{M}_i \cdot \mathrm{diag}([z_{i,j}]_{j=1}^{K_i})\\ z_{i,j} &\sim Bernoulli(p_i), i = 1, 2, \ldots L, j = 1, 2, \ldots K_{i-1}\end{aligned} \tag{8}$$

where $z_{i,j}$ denotes the binary variable obtained from the Bernoulli distribution given probability p_i. If $z_{i,j} = 0$, the j-th unit of the $(i-1)$-th layer is dropped out. When performing inference during execution, we also enable dropout in our Bayesian deep graph matching approach to sample $\mathbf{W}$. That is, the distribution of correspondence is inferred by:

$$p(\mathbf{Y}|\mathcal{G},\mathcal{G}',\mathcal{X},\mathcal{Y}) \approx \frac{1}{T}\sum_{t=1}^{T} p(\mathbf{Y}|\mathcal{G},\mathcal{G}',\mathbf{W}^{(t)}), \mathbf{W}^{(t)} \sim q(\mathbf{W}) \tag{9}$$

where T is the number of sampling. We define the final correspondence as the expectation of the correspondence samples sampled from Eq. (9), which is denoted as $\mathbb{E}(p(\mathbf{Y}))$, where $\mathbb{E}$ denotes the expectation function. The uncertainty of each correspondence is defined as follows:

$$\mathbb{H}(\mathbb{E}(p(\mathbf{Y})_{i,j})) = -\mathbb{E}(p(\mathbf{Y}_{i,j})) * \log(\mathbb{E}(p(\mathbf{Y}_{i,j}))) \tag{10}$$

where $\mathbb{H}$ is the Shannon entropy. The entropy encodes the total uncertainty in the correspondence results including both data uncertainty in robot observations and model uncertainty in the graph network (Depeweg et al., 2017).

The loss function for our Bayesian deep graph matching approach is defined as follows:

$$\mathcal{L}_{coid} = -\log\left(\frac{1}{nn'}||\mathbf{S} \circ \mathbf{Y}^* \circ \mathbb{E}(\mathbf{Y})||_1\right) \tag{11}$$

where $\circ$ represents the element-wise product, n and n' are the number of objects in graph $\mathcal{G}$ and $\mathcal{G}'$ respectively, and $\mathbf{Y}^*$ denotes the ground truth of the correspondence matrix, with $\mathbf{Y}^*_{i,i'} = 1$ denoting the ground truth of correspondence between the i-th object in graph $\mathcal{G}$ and the i'-th object in graph $\mathcal{G}'$. Because the negative log loss requires the value in range of $[0, 1]$, we use sum-averaged function to normalize the overall similarity. Given the Bayesian dropout approximation theory (Gal & Ghahramani, 2016), minimizing the negative-log loss function $\mathcal{L}_{coid}$ is equivalent to the minimization of the KL-divergence in Eq. (7). Accordingly, training our proposed deep graph matching model with gradient descent enables the learning of an approximated distribution of weights, which allows us to quantify uncertainty in the identified correspondence results.

3.3 Reducing perceptual non-covisibility and correspondence uncertainty

Since non-covisible objects are observed only by one robot, they do not have correspondences. To explicitly address this challenge, we design a novel loss function that integrates non-covisibility into the learning process, which is defined as follows:

$$\mathcal{L}_{non} = -\log\left(\frac{1}{nn'}||\exp(-\mathbf{S} \circ \mathbf{N} \circ \mathbb{E}(\mathbf{Y}))||_1\right) \tag{12}$$

where $\exp(-)$ denotes an element-wise negative exponential operator, which is used to normalize the penalty caused non-covisibility between $[0, 1]$. Given $\exp(-)$, minimizing the loss $\mathcal{L}_{non}$ is equivalent to minimizing the non-covisibility penalty. $\mathbf{N} \in \mathcal{R}^{n\times n'}$ denotes an indicator matrix that includes the indices of non-covisible objects in $\mathbf{Y}$, with $\mathbf{N}_{i,i'} = 1$ indicating that the correspondence $\mathbf{Y}_{i,i'}$ is constructed by non-covisible objects. For example, if the i-th object in graph $\mathcal{G}$ or the i'-th object in graph $\mathcal{G}'$ is non-covisible object which has no correspondence, then $\mathbf{N}_{i,i'} = 1$. In Eq. (12), we first calculate the similarity of the correspondences constructed by non-covisible objects as $\mathbf{S} \circ \mathbf{N} \circ \mathbb{E}(\mathbf{Y})$. Then, the similarity of non-covisible objects is converted to a normalized penalty term and added to the overall loss.

Similarly, we also explicitly model the quantified uncertainty as a penalty term that is added to $\mathcal{L}_{coid}$ to improve the robustness of deep graph matching, which is defined as:

$$\mathcal{L}_{unc} = -\log\left(\frac{1}{nn'}||\exp\left(-\mathbb{H}\left(\mathbb{E}(\mathbf{Y})\right)\right)||_1\right) \tag{13}$$

where $\mathbb{H}(\mathbb{E}(\mathbf{Y}))$ is our quantified uncertainty in the identified correspondences.

Our final loss function is represented as $\mathcal{L} = \mathcal{L}_{coid} + \mathcal{L}_{non} + \mathcal{L}_{unc}$. Minimizing this loss function during training is equivalent to maximizing the similarity of correct correspondences and minimizing the similarity of non-covisibile objects and matching uncertainty. During execution, given the quantified uncertainty in the identified correspondence, we further improve the correspondences results by defining a threshold λ, in order to remove the correspondences with high uncertainty values (Gao et al., 2020a). Specifically, if $\mathbb{H}(\mathbb{E}(p(\mathbf{Y})_{i,i'})) \geq \lambda$, the correspondence $\mathbf{Y}_{i,i'}$ is removed.

3.4 Multi-robot collaborative object localization based on identified correspondences

Object localization is an important research topic in robotics and computer vision, with the goal of improving situational awareness for robots. Existing techniques for object localization is mainly focusing on single-robot multi-sensory data fusion, e.g., fusing position and velocity information based on Kalman filters (Weng et al., 2020) and integrating Kinematic and RGB measurements for pose estimation (Qin et al., 2019). However, none of these methods consider multi-robot sensor fusion, which is much more challenging than single-robot multi-sensory data fusion due to the dynamics of robots (Queralta et al., 2020).

Recently, collaborative object localization attracts more and more attention, which uses a team of robots to perform object localization by integrating multi-robot observations of objects, in order to improve object localization accuracy and resilience to sensor failures (Gao et al., 2021a, b). To perform collaborative object localization, identifying correspondence of objects in multi-robot observations is required.

In this paper, we propose a principled method to collaboratively localize objects based on identified correspondences, in order to improve the accuracy of the measured locations of objects. Specifically, given a multi-robot system with N robots ($N \geq 2$). For each pair of observations provided by a pair of robots, the pairwise correspondences of objects (observed by both of the robots) can be founded via our proposed correspondence identification approach as defined in Eqs. (1–4). Given the pairwise correspondences, we can identify the covisible object among all N robots' observations by forcing circle consistency (Fathian et al., 2020). We define the locations of identified objects as $\{\mathbf{v}_1^n, \mathbf{v}_2^n, \ldots, \mathbf{v}_M^n\}, n = 1, 2, \ldots, N$, where $\mathbf{v}_i^n$ denotes the measured location of the i-th object detected by the n-th robot and M denotes the number of covisible objects observed by the N robots. M and N can be variant. We assume that all robots' measurements are independent.

For simplification, we use $\mathbf{v}^i$ denotes the measured location acquired by the i-th robot in the follows. Since the number of robots to collaborate with is arbitrary, we propose a multi-robot fusion gain to integrate arbitrary number multi-robot measurements, which is defined as follows:

$$\mathbf{M}^i = \left(\sum_{j=1}^{N}(\mathbf{P}^j)^{-1}\right)^{-1}(\mathbf{P}^i)^{-1} \tag{14}$$

where $\mathbf{M}^i \in \mathcal{R}^{3\times3}$ denotes the measurement fusion gain for the i-th robot. In addition, $\mathbf{M}^i$ follows the constraint $\sum_{i=1}^{N}\mathbf{M}^i = \mathbf{I}$, where $\mathbf{I} \in \mathcal{R}^{3\times3}$ denotes an identity matrix. The fusion gain represents the weight of each robot's measurement in all the multi-robot measurements given the normalized location uncertainties. For each single robot, its final estimation of an object's location is computed via integrating its own estimation and its collaborators' estimations weighted by these fusion gains. Formally, it is defined as follows:

$$\hat{\mathbf{v}}^i = \mathbf{M}^i\mathbf{v}^i + \sum_{j=1, j\neq n}^{N}\mathbf{M}^j\sigma(\mathbf{v}^j) \tag{15}$$

where σ denotes a transformation function that transforms the locations of objects from the j-th robot's coordinates to the i-th robot's coordinates, which is computed through using camera extrinsic parameters (Zhang & Pless, 2004). The camera extrinsic parameters can be obtained through GPS or deep learning algorithms (Kendall et al., 2015). In addition, $\hat{\mathbf{v}}^i$ denotes the final location estimation of the target object observed by the i-th robot. The final location is computed by the sum of multi-robot measurements $\mathbf{v}^j, i = 1, 2, \ldots, N$, weighted by the fusion gains $\mathbf{M}^j$. If a robot's measurements of objects' locations have large noise, then its contribution will be heavily weakened during the fusion. The uncertainty of the final location estimation is defined as follows:

$$\hat{\mathbf{P}}^i = \left(\sum_{n=1}^{N}(\mathbf{P}^i)^{-1}\right)^{-1} \tag{16}$$

where $\hat{\mathbf{P}}^i$ denotes the uncertainty of the final location estimation $\hat{\mathbf{v}}^i$, which is obtained by integrating all the multi-robot location uncertainties. It is worth noting that the updated state

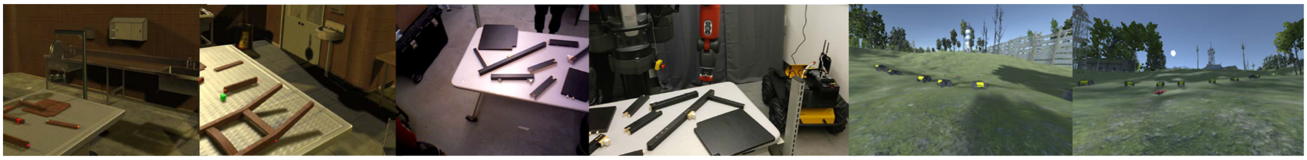

(a) Observations by two robots in SFAT (b) Observations by a robot and a human wearing an AR headset in RFAT (c) Observations by two robots in SMRC

Fig. 2 Examples of the color image observations that are acquired by a pair of agents from different perspectives in the experimental scenarios of SFAT, RFAT and SMRC

estimations and uncertainties can be further improved when new measurements become available (Pei et al., 2019). Thus, the fusion process can be run incrementally.

4 Experiments

We evaluate our approach with simulations and physical robots in three scenarios. Specifically, we examine the experimental results of our approach compared with previous methods and discuss the characteristics of our approach.

4.1 Experimental setups

We use two high-fidelity robotics simulations and physical robots to evaluate our method for correspondence identification in collaborative perception applications, including Simulated furniture assembly tasks (SFAT) as shown in Fig. 2a, Real-world furniture assembly tasks (RFAT) as shown in Fig. 2b and Simulated multi-robot coordination (SMRC) as shown in Fig. 2c.

We construct each observation as a graph with node attributes generated from appearance features (Gao et al., 2020a). The edges are generated by Delaunay triangulation given the 2D camera coordinates of objects in SFAT and RFAT and 3D real world coordinates of objects in SMRC. For the B-Spline GNN Ψ, we set the number of convolutional layers $L = 2$ with each layer using a kernel size of 5 in each dimension and a hidden dimensionality of 256. Each convolutional layer is followed by dropout with probability 0.4. For the MLP φ, each linear layer is followed by dropout with probability 0.2. In all the experiments, we use ADMM as the optimization method. We run 150, 250, 100 epochs for our approach in SFAT, RFAT and SMRC, respectively. The number of samplings T for Bayesian inference is set to 20.

We implement the full version of our approach using $\mathcal{L} = \mathcal{L}_{coid} + \mathcal{L}_{non} + \mathcal{L}_{unc}$ as the loss function. We also implement two baseline methods, using $\mathcal{L}_{coid+non} = \mathcal{L}_{coid} + \mathcal{L}_{non}$ that addresses only non-covisibility, and $\mathcal{L}_{coid+unc} = \mathcal{L}_{coid} + \mathcal{L}_{unc}$ that addresses only uncertainty. In addition, we compare our approach with four previous correspondence identification methods, including two learning-free graph matching methods and two deep learning-based methods. They are:

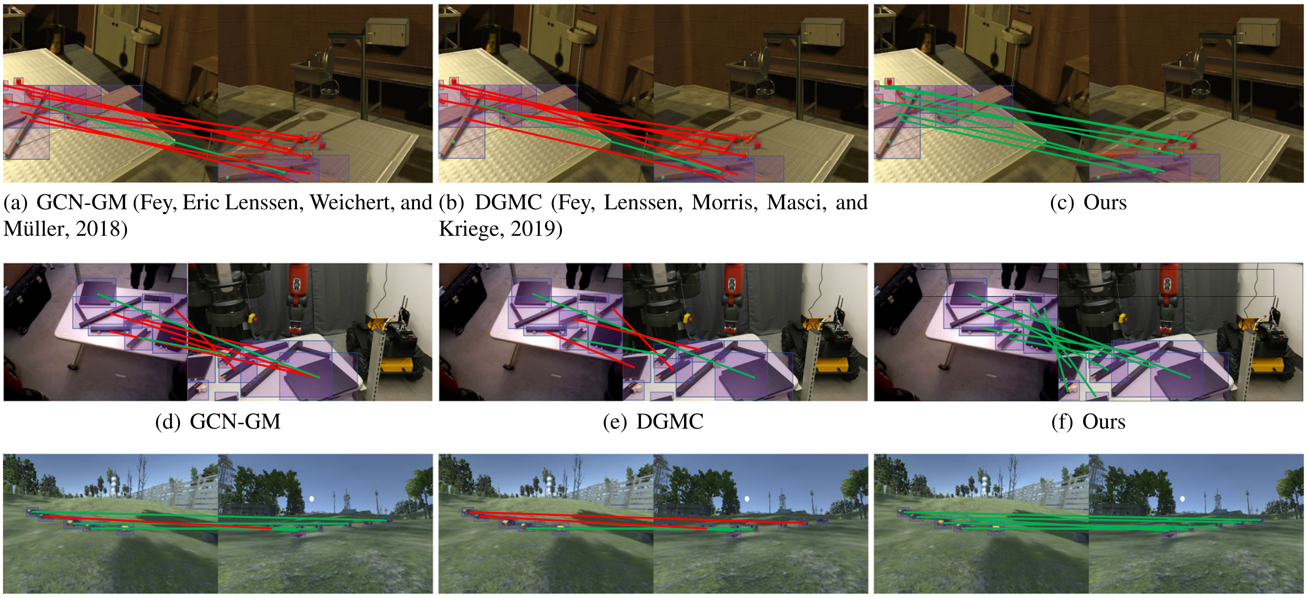
- Multi-order graph matching (**MOGM**) (Chang et al., 2017), which integrates multiple different attributes in a learning-free way to identify correspondences.
- Regularized graph matching (**RGM**) (Gao et al., 2020a), which addresses perception uncertainty and non-covisible objects in a learning-free way to identify correspondences.
- Graph convolutional network-based graph matching (**GCN-GM**) (Fey et al., 2018), which identifies correspondences by only optimizing the loss of overall similarity between two observations.
- Deep graph matching consensus (**DGMC**) (Fey et al., 2019), which uses the similarity of embedding vectors obtained by graph neural networks for correspondence identification while checking the neighborhood consensus of identified correspondences.

Following a standard experimental setup (Cho et al., 2010; Gao et al., 2020a), precision and recall are adopted to evaluate our approach. Given the identified correspondences, precision is defined as the ratio of correct correspondences over all the identified correspondences. Recall is defined as the ratio of identified correspondences over all ground truth correspondences. In addition, we also use F1 score as a measurement of the overall performance, which is defined as $\frac{2pr}{(p+r)}$, where p denotes the precision and r denotes the recall.

4.2 Results on furniture assembly simulations

Our approach is first evaluated on SFAT, in which the correspondences of objects are identified for multi-robot collaborative furniture assembly. Correspondence identification is used to make the robots refer to the same object in their respective field of view. The SFAT scenario is challenging due to the existence of a large number of non-covisible objects and strong occlusion in multi-robot observations.

SFAT consists of three subtasks, including assembling a shelf, chair and table. Each subtask includes 750 data instances. Each instance consists of a pair of RGB images



(a) GCN-GM (Fey, Eric Lenssen, Weichert, and Müller, 2018)

(b) DGMC (Fey, Lenssen, Morris, Masci, and Kriege, 2019)

(c) Ours

(d) GCN-GM

(e) DGMC

(f) Ours

(g) GCN-GM

(h) DGMC

(i) Ours

Fig. 3 Qualitative experimental results of our approach over SFAT (first row), RFAT (second row), and SMRC (third row), and comparisons with GCN-GM and DGMC. Green lines denote correct correspondences and red lines denote incorrect correspondences. [Best viewed in color.]

Table 1 Quantitative results based on the metrics of precision and recall over SFAT, RFAT and SMRC

Method	SFAT		RFAT		SMRC	
	Recall	Precision	Recall	Precision	Recall	Precision
MOGM	0.4385	0.2332	0.2298	0.2467	0.7184	0.7136
RGM	0.4434	0.2841	0.2871	0.3012	0.7878	0.7735
GCN-GM	0.9078	0.5398	0.7580	0.8916	0.9321	0.8481
DGMC	0.9105	0.5441	0.9933	0.8971	0.9388	0.9037
$\mathcal{L}_{coid+non}$	0.9122	0.5526	**0.9960**	0.9036	0.9477	0.9319
$\mathcal{L}_{coid+unc}$	0.9053	0.7011	0.9937	0.9038	**0.9529**	0.9611
Ours	**0.9216**	**0.7026**	0.9920	**0.9498**	0.9503	**0.9683**

Bold values indicate the best performance

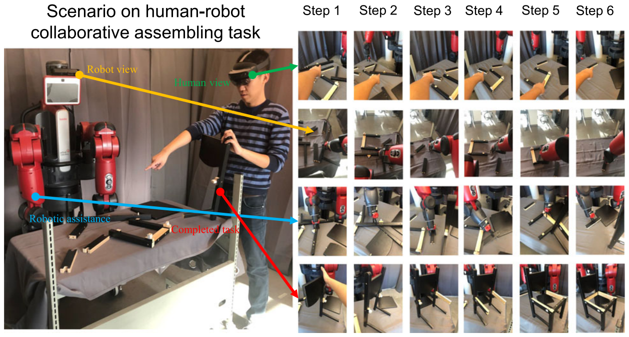


Fig. 4 Illustrations of several steps in the scenario of robot-assisted furniture assembly. The Baxter robot assists a human collaborator who wears an AR headset to collaboratively assemble an IKEA chair

observed by two robots from different perspectives. In each image, at least 5 objects are detected. The ground truth correspondences are obtained from the simulator (Lee et al., 2019). 400 data instances are used for training and 350 instances are used for testing. The quantitative results are obtained by averaging 4 times of the experiments.

The qualitative results obtained by our approach on SFAT are presented in Fig. 3c. We can see that our approach can accurately identify correspondences. Compared with GCN-GM and DGMC as shown in Fig. 3a and b, our approach obtains a significant improvement when faced with strong non-covisibility and perception uncertainty caused by occlusion. In addition, our method can remove correspondences with highly quantified uncertainty, which can further reduce the number of incorrect correspondences caused by this uncertainty and non-covisibility.

The quantitative results from SFAT are presented in Table 1. We observe that our baseline methods $\mathcal{L}_{coid} + \mathcal{L}_{non}$ and $\mathcal{L}_{coid} + \mathcal{L}_{unc}$ generally achieve better performance than the deep-learning methods GCN-GM and DGMC, as GCN-GM and DGMC only focus on minimizing the loss of the overall similarity. Thus, the results indicate the importance of addressing non-covisibility and correspondence uncertainty in correspondence identification. Since only 2D spatial information is available in SFAT, learning-free methods MOGM and RGM perform poorly due to their reliance on high-quality observations. The deep learning-based methods GCM-GM and DGMC perform significantly better due to their learning capability. The full version of our approach obtains the best performance due to its ability to address non-covisibility and perception uncertainty in multi-robot assembly tasks.

4.3 Results in real-world furniture assembly scenarios

Our approach is further evaluated on RFAT, in which a human and a robot collaboratively assembly an IKEA chair. Figure 4 provides the details of the scenario, in which the Baxter robot assists a human collaborator wearing an AR headset to assemble an IKEA chair. The RFAT scenario is challenging as it contains a diverse set of furniture parts observed by the robot and the human collaborator from two different perspectives and both of the perspectives contain a large number of non-covisible objects and strong occlusion in the observations.

RFAT includes 500 data instances. Each instance includes a pair of RGB images obtained by a robot and a human who wears a Hololen2 AR headset. In each image, at least 5 objects are detected. The ground truth correspondences are obtained through the Scalabel software (Yu et al., 2020). 250 data instances are used for training and 250 instances are used for testing.

The qualitative results obtained by our approach in RFAT are presented in Fig. 3f. We can observe that our approach can accurately identify correspondences and obtain a significant improvement over the other graph learning methods (GCN-GM and DGMC). In this scenario, the existence of strong non-covisibility and perception uncertainty hinders the performance of deep learning-based methods GCN-GM and DGMC, which only minimize the similarity loss during learning. Our approach can address these challenges by integrating non-covisibility and perception uncertainty into the learning process. By quantifying uncertainties of correspondences, our method can further reduce the number of incorrect correspondences caused by perception uncertainty and non-covisibility.

The quantitative results obtained in RFAT are presented in Table 1. We can see that our baseline methods $\mathcal{L}_{coid} + \mathcal{L}_{non}$ and $\mathcal{L}_{coid} + \mathcal{L}_{unc}$ outperform the deep learning-based methods GCN-GM and DGMC, which only consider minimizing the loss on the overall similarity. Our full version approach obtains the best performance (based on the F1 score) by addressing non-covisibility and perception uncertainty for correspondence identification in human-robot collaborative assembly task.

Table 2 Quantitative analysis on the influence of thresholding the identified correspondences based on the quantified uncertainty. The metric reported is the F1-score over SFAT, RFAT and SMRC

Method	Before threshold	After threshold
SFAT	0.7009	**0.8303**
RFAT	0.9695	**0.9724**
SMRC	0.9456	**0.9686**

Bold values indicate the best performance

4.4 Results in multi-robot coordination scenarios

Our approach is finally evaluated in the scenario of multi-robot coordination, in which a group of robots is observed by two ground robots. In the observations, there exists strong perception uncertainty caused by long distances between the observers and the observed objects, low resolution of the acquired images, and the lack of textures of objects in observations.

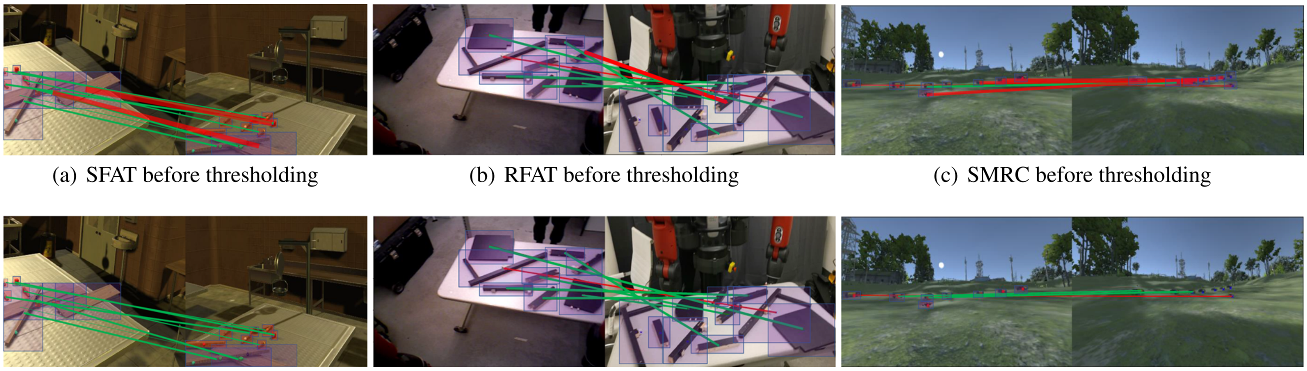
SMRC includes 600 data instances. Each instance is recorded by two robots from different perspectives and includes a pair of RGB images with at least 7 detected objects, with depth images and ground truth correspondences obtained from the simulation. We use 200 instances for training and 400 instances for testing.

The qualitative results of our approach in SMRC are shown in Fig. 3i. We observe that our approach can correctly identify the correspondences. The results of GCN-GM and DGMC are shown in Fig. 3g and h separately. It is observed that the objects far away from the camera are identified incorrectly due the perception uncertainty caused by the low resolution of objects. In addition, GCN-GM and DGMC focus on maximizing the overall similarity, which is affected by non-covisibility. Thus, addressing correspondence uncertainty and non-covisibility are important for correspondence identification.

The quantitative results on SMRC are presented in Table 1. Due to the 3D information provided by SMRC, MOGM and RGM obtain superior results compared to their results in SFAT and RFAT. The deep learning-based methods GCN-GM and DGMC further improve on this performance due to their learning capability. Our approach achieves the best performance compared with these four methods by addressing non-covisibility and perception uncertainty in the multi-robot coordination scenario.

4.5 A case study in connected autonomous driving

As a case study, we finally deploy our proposed approach in the connected autonomous driving (CAD) scenario, in which



(a) SFAT before thresholding

(b) RFAT before thresholding

(c) SMRC before thresholding

(d) SFAT after thresholding

(e) RFAT after thresholding

(f) SMRC after thresholding

Fig. 5 Qualitative experimental results of our approach, with identified correspondences thresholded based upon the quantified correspondence uncertainties. Green lines denote correct correspondences and red lines denote incorrect correspondences. A wider line denotes a greater value of uncertainty in the identified correspondence. [Best viewed in color.]

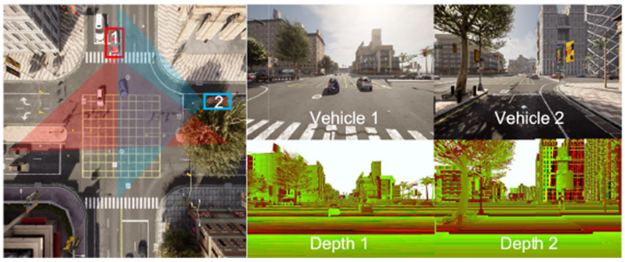


Fig. 6 Illustrations of the connected autonomous driving scenario

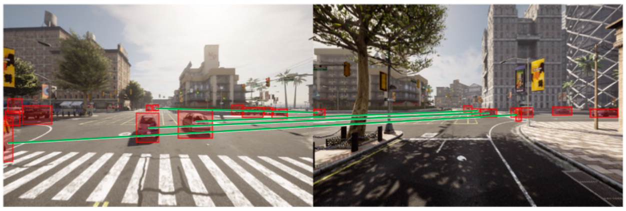


Fig. 7 A qualitative result obtained by our approach perform in the CAD scenario

two connected vehicles collaboratively localize the street objects (e.g., vehicles and pedestrians), in order to improve each others' performance of object localization. Figure 6 provides the details of the scenario, in which two connected vehicles meet at a street intersection, they want to collaborate with each other to improve the localization accuracy of the street objects. The CAD scenario is challenging as it contains various kinds of objects on the street. These street objects are highly dynamic, strongly occluded or have low resolution due to the long-distance observations.

The CAD scenario is conducted in a high-fidelity simulation, which is implemented by CARLA (Dosovitskiy et al., 2017), with the traffic trajectories designed by SUMO (Behrisch et al., 2011). CAD includes 300 data instances. Each instance includes a pair of RGB images obtained by two connected vehicles. In each image, at least one object can be observed by both of the connected vehicles. The object detection is performed by Yolo v5 (Jocher, 2020). The ground truth correspondences are obtained through the unique IDs of objects provided by the simulator. 200 data instances are used for training and 100 instances are used for testing.

The qualitative results of our approach are presented in Fig. 7. Given the results, we can see that our approach can well identify the correspondences of the street objects between the observations acquired by connected vehicles. From the quantitative perspective, our approach achieves 0.8513 precision and 0.6271 recall in this challenging scenario with large number of non-covisible and highly dynamic objects, as well as strong perception uncertainty.

Based on the identified correspondences, we further evaluate our proposed approach of multi-robot collaborative object localization. We use displacement error as the metrics, which is defined as the Euclidean distance between our estimated locations and the ground truth locations. Then, we compared our proposed approach with the baseline method that only uses single-robot object location measurements. In the location measurements provided by the depth camera, we add a Gaussian noise to it following the recent collaborative object localization approach (Gao et al., 2021a). Based upon the evaluation metrics, the baseline method obtains 1.9961m displacement error and our approach obtains 1.2548m displacement error, which indicate the significant improvement achieved by our approach.

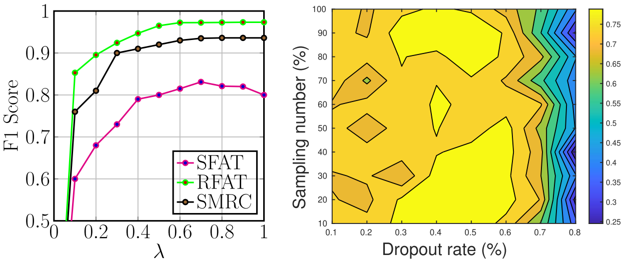
4.6 Discussion

We further evaluate various characteristics of our approach, including the importance of uncertainty quantification in cor-

Table 3 Quantitative analysis on the performance of our approach using epistemic, aleatoric, and Shannon entropy uncertainty (Depeweg et al., 2017). The metric reported is the F1-score over SFAT, RFAT and SMRC

Methods	SFAT	RFAT	SMRC
Epistemic	0.7009	0.9722	0.9456
Aleatoric	**0.8303**	0.9695	0.9676
Shannon Entropy	0.8143	**0.9724**	**0.9688**

Bold values indicate the best performance



(a) Threshold effects

(b) Dropout rate and sampling number

Fig. 8 Hyperparameter analysis based on the metric of F1 scores

respondence identification, the performance of our approach using different uncertainties, and hyperparameter analysis.

4.6.1 Uncertainty quantification in correspondence identification

Figure 5 shows the effect of quantifying the correspondence uncertainty on correspondence identification. We can see that incorrect correspondences correspond to objects with large perception uncertainty caused by occlusion, which leads to a much larger correspondence uncertainty for incorrect correspondences (visualized with a red line, with the width representing uncertainty) than the correct correspondences (visualized with a green line). Given the quantified correspondence uncertainty, we can further improve the correspondences results by defining a threshold λ, in order to remove the correspondences with high uncertainty values. As shown in Table 2, the performance of our approach in all three scenarios is improved by thresholding the correspondences given the quantified uncertainties. Thus, utilizing the quantified uncertainty for correspondence identification can effectively reduce the number of incorrect correspondences.

4.6.2 Different types of uncertainties

One of our proposed novelties is to integrate the quantified uncertainty into the loss function and to use it for the removal of incorrect correspondences. Thus, we analyze the performance of our approach by using three different types of uncertainty for correspondence identification, including epistemic uncertainty, aleatoric uncertainty, and the Shannon entropy (the sum of epistemic and aleatoric uncertainty). Epistemic uncertainty is defined as the ambiguity in the learning model (e.g. caused by the out-of-distribution data) and aleatoric uncertainty represents the ambiguity of data (e.g. caused by low texture regions in observations) (Depeweg et al., 2017). Shannon entropy represents the total uncertainty, as defined in Eq. (10). Given the F1 scores reported in Table 3, we can see that using aleotoric uncertainty achieves the best performance in SFAT, which indicates the presence of large data uncertainty caused by perception uncertainty in this scenario. The poor performance obtained from using epistemic uncertainty indicates the low model uncertainty in SFAT due to the large amount of training data. In RFAT and SMRC, the improved performance obtained from using epistemic uncertainty indicates large uncertainty in the learning model. Shannon entropy generally performs the best due to the representation of both model and data uncertainty.

4.6.3 Hyperparameter analysis

We use the hyperparameter λ to threshold the identified correspondences based on the quantified correspondence identification, in order to remove incorrect correspondences with high uncertainty. We randomly choose 80 pairs of graphs in each of SFAT, RFAT and SMRC, and perform sensitivity analysis to analyze the performance influenced by λ based on the F1 score. As shown in Fig. 8a, the results indicate that our approach obtains the best performance when $\lambda = 0.7$ on different scenarios.

The performance of our approach is also influenced by the dropout rate and sampling numbers of our model. Based on the F1 score, we evaluate the performance of our approach in the SFAT scenario with the dropout rate in the range of [0.1, 0.8] and the sampling number in the range of [10, 100]. Given the results shown in Fig. 8b, we can see that our approach obtains the best performance when the dropout rate is in the range of [0.4, 0.5] and the performance decreases fast as the dropout rate increases from 0.6 to 0.8. The sampling number has several optimal values in our evaluation range, including [20, 30], [50, 60] or [80, 90].

5 Conclusion

It is important to address correspondence identification in order to enable multiple agents (including robots and humans) to refer to the same objects within their own fields of view when performing collaborative tasks. To address the key shortcomings of the current deep graph matching methods, including the lack of ability to reduce correspondence uncertainty and perceptual non-covisibility, we propose a novel method using Bayesian deep graph matching for correspondence identification. Our method formulates correspondence identification in collaborative perception as a deep graph matching problem under the Bayesian learning framework to quantify correspondence uncertainty. We improve our approach's robustness by explicitly penalizing correspondences with high uncertainty values and correspondences caused by non-covisible objects. Based on the identified correspondences, we further design a new approach to perform multi-robot collaborative object localization, which improves the object localization accuracy by integrating multi-robot observations. Extensive experiments are conducted to evaluate our method in collaborative furniture assembly, multi-robot coordination and connected autonomous driving applications based on high-fidelity simulations and physical robots. Experimental results show that our method outperforms the previous and baseline methods and achieves state-of-the-art performance of correspondence identification in collaborative perception.

Even though our approach achieves the state-of-the-art performance, we would like to improve our current approach from the following directions: (1) For the outdoor connected autonomous driving scenario, besides RGB-D data, we will also consider multi-sensory data, e.g., GPS, and IMU, in order to improve the performance of correspondence identification; (2) The current collaborative object localization is only performed to the objects that can be observed by all the robots, we will further study the cases with missing objects in a observation sequence; (3) The current correspondence identification is performed between a pair of robots, we will further study the correspondence identification among more than two robots' observations.

Acknowledgements This work was partially supported by NSF CAREER Award IIS-1942056 and DARPA YFA D21AP10114-00.

Declaration

Conflict of interest The conflict of interest of this work includes the Colorado School of Mines (@mines.edu), the University of Massachusetts Amherst (@cs.umass.edu), and the University of Maryland, College Park (@umd.edu).

References

Acevedo, J. J., Messias, J., Capitán, J., Ventura, R., Merino, L., & Lima, P. U. (2020). A dynamic weighted area assignment based on a particle filter for active cooperative perception. *IEEE Robotics and Automation Letters, 5*(2), 736–743.

Behrisch, M., Bieker, L., Erdmann, J., & Krajzewicz, D. (2011). Sumo–simulation of urban mobility: An overview. In *Proceedings of SIMUL 2011,* The Third International Conference on Advances in System Simulation.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In *International Conference on Machine Learning,* pp. 1613–1622

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3(1):1–122.

Brambilla, M., Ferrante, E., Birattari, M., & Dorigo, M. (2013). Swarm robotics: A review from the swarm engineering perspective. *Swarm Intelligence, 7*(1), 1–41.

Chang, H. J., Fischer, T., Petit, M., Zambelli, M., & Demiris, Y. (2017). Learning kinematic structure correspondences using multi-order similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 2920–2934.

Cho M, Lee J, Lee KM (2010) Reweighted random walks for graph matching. In *European Conference on Computer Vision*.

Chung, S. J., Paranjape, A. A., Dames, P., Shen, S., & Kumar, V. (2018). A survey on aerial swarm robotics. *IEEE Transactions on Robotics, 34*(4), 837–855.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., Udluft, S. (2017). Uncertainty decomposition in bayesian neural networks with latent variables. arXiv preprint.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning,* pp. 1–16.

Engel J, Schöps, T., Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*.

Fathian, K., Khosoussi, K., Tian, Y., Lusk, P., & How, J. P. (2020). CLEAR: A consistent lifting, rmbedding, and alignment rectification algorithm for multi-agent data association. *IEEE Transactions on Robotics, 36*(6), 1686–1703.

Fey, M., Eric Lenssen, J., Weichert, F., Müller, H. (2018) SplineCNN: Fast geometric deep learning with continuous b-spline kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fey, M., Lenssen, J. E., Morris, C., Masci, J., Kriege, N. M. (2019). Deep Graph Matching Consensus. In *International Conference on Learning Representations*.

Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. arXiv.

Frey, K. M., Steiner, T. J., & How, J. P. (2019). Efficient constellation-based map-merging for semantic SLAM. In *IEEE International Conference on Robotics and Automation*.

Gal, Y., & Ghahramani, Z. (2015). Bayesian convolutional neural networks with Bernoulli approximate variational inference.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*.

Gao, P., & Zhang, H. (2021). Bayesian deep graph matching for correspondence identification in collaborative perception. *Robotics: Science and Systems*.

Gao, P., Guo, R., Lu, H., & Zhang, H. (2020a). Regularized graph matching for correspondence identification under uncertainty in collaborative perception. *Robotics: Science and Systems*.

Gao, P., Reily, B., Paul, S., Zhang, H. (2020b). Visual reference of ambiguous objects for augmented reality-powered human-robot communication in a shared workspace. In *International Conference on Human-Computer Interaction*.

Gao, P., Guo, R., Lu, H., & Zhang, H. (2021a) Multi-view sensor fusion by integrating model-based estimation and graph learning for collaborative object localization. *ICRA*.

Gao, P., Reily, B., Guo, R., Lu, H., Zhu, Q., & Zhang, H. (2021b). Asynchronous collaborative localization by integrating spatiotemporal graph learning with model-based estimation. arXiv preprint.

Guo, R., Lu, H., Gao, P., Zhang, Z., & Zhang, H. (2019). Collaborative localization for occluded objects in connected vehicular platform. In *IEEE 90th Vehicular Technology Conference*.

Hietanen, A., Pieters, R., Lanz, M., Latokartano, J., & Kämäräinen, J. K. (2020). Ar-based interaction for human-robot collaborative manufacturing. *Robotics and Computer-Integrated Manufacturing, 63*, 101891.

Jiang, B., Sun, P., Tang, J., & Luo, B. (2019). Glmnet: Graph learning-matching networks for feature matching. arXiv.

Jin, X., Lan, C., Zeng, W., Wei, G., & Chen, Z. (2020). Semantics-aligned representation learning for person re-identification. In *AAAI Conference on Artificial Intelligence*.

Jocher, G. (2020). ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements.

Kendall, A, & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *International Conference on Neural Information Processing Systems*.

Kendall, A., Grimes, M., & Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *IEEE International Conference on Computer Vision*.

Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Khatun, A., Denman, S., Sridharan, S., & Fookes, C. (2020). Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision*.

Kupinski, M. A., Hoppin, J. W., Clarkson, E., & Barrett, H. H. (2003). Ideal-observer computation in medical imaging with use of Markov-chain Monte Carlo techniques. *Journal of the Optical Society of America A, 20*(3), 430–438.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *International Conference on Neural Information Processing Systems*.

Lee, Y., Hu, E. S., Yang, Z., Yin, A., & Lim, J. J. (2019). IKEA furniture assembly environment for long-horizon complex manipulation tasks. arXiv.

Leordeanu, M., Hebert, M. (2005). A spectral technique for correspondence problems using pairwise constraints. In *IEEE International Conference on Computer Vision*.

Lou, Z., You, J., Wen, C., Canedo, A., & Leskovec, J., et al. (2020). Neural subgraph matching. arXiv.

Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. In *International Conference on Neural Information Processing Systems*.

Maset, E., Arrigoni, F., & Fusiello, A. (2017). Practical and efficient multi-view matching. In *IEEE International Conference on Computer Vision*.

Matsumoto, S., & Riek, L. D. (2019). Fluent coordination in proximate human robot teaming. In *Robotics: Science and Systems workshop*.

Nguyen, Q., Gautier, A., & Hein, M. (2015). A flexible tensor block coordinate ascent scheme for hypergraph matching. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Nie, W.Z., Liu, A.A., Gao, Z., & Su, Y.T. (2015). Clique-graph matching by preserving global & local structure. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Pei, Y., Biswas, S., Fussell, D. S., & Pingali, K. (2019). An elementary introduction to kalman filtering. *Communications of the ACM, 62*(11), 122–133.

Qin F., Li, Y., Su, Y.H., Xu, D., & Hannaford, B. (2019). Surgical instrument segmentation for endoscopic vision with data fusion of rediction and kinematic pose. In *ICRA*.

Queralta, J. P., Taipalmaa, J., Pullinen, B. C., Sarker, V. K., Gia, T. N., Tenhunen, H., Gabbouj, M., Raitoharju, J., & Westerlund, T. (2020). Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision. *IEEE Access, 8*, 191617–191643.

Quispe, R., & Pedrini, H. (2019). Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing, 92*, 103809.

Reily, B., Han, F., Parker, L. E., & Zhang, H. (2018). Skeleton-based bio-inspired human activity prediction for real-time human-robot interaction. *Autonomous Robots, 42*(6), 1281–1298.

Reily, B., Reardon, C., & Zhang, H. (2020). Representing multi-robot structure through multi-modal graph embedding for the selection of robot teams. arXiv.

Rolínek, M., Swoboda, P., Zietlow, D., Paulus, A., Musil, V., & Martius, G. (2020). Deep graph matching via blackbox differentiation of combinatorial solvers. *European Conference on Computer Vision*.

Ryu, S., Kwon, Y., & Kim, W. Y. (2019). A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chemical Science, 10*(36), 8438–8446.

Shi H, Yang Y, Zhu X, Liao S, Lei Z, Zheng W, Li SZ (2016) Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*.

Suh Y, Adamczewski K, Mu Lee K (2015) Subgraph matching using compactness prior for robust feature correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Tian, Y., Liu, K., Ok, K., Tran, L., Allen, D., Roy, N., How, J. P. (2019). Search and rescue under the forest canopy using multiple UAVs. *The International Journal of Robotics Research*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., & Leibe, B. (2019). MOTS: Multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. (2019a). Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, R., Yan, J., & Yang, X. (2019b). Learning combinatorial embedding networks for deep graph matching. In *IEEE International Conference on Computer Vision*.

Wei, S., Yu, D., Guo, C. L., Dan, L., & Shu, W. W. (2018). Survey of connected automated vehicle perception mode: From autonomy to interaction. *Intelligent Transport Systems, 13*(3), 495–505.

Weng, X., Wang, J., Held, D., & Kitani, K. (2020). 3D multi-object tracking: A baseline and new evaluation metrics. *IROS*.

Yan, J., Ren, Z., Zha, H., & Chu, S. (2016). A constrained clustering based approach for matching a collection of feature sets. In *International Conference on Pattern Recognition*.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conference on Computer Vision and Pattern Recognition*

Yu, H. X., Wu, A., & Zheng, W. S. (2018). Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yue, Y., Zhao, C., Li, R., Yang, C., Zhang, J., Wen, M., Wang, Y., Wang, D. (2020). A hierarchical framework for collaborative probabilistic semantic mapping. In *IEEE international conference on robotics and automation*.

Zhang, B., Choudhury, S., Hasan, M. A., Ning, X., Agarwal, K., Purohit, S., & Cabrera, P. P. (2016). Trust from the past: Bayesian personalized ranking based link prediction in knowledge graphs. arXiv.

Zhang, Q., & Pless, R. (2004). Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Zhang, Y., Pal, S., Coates, M., & Ustebay, D. (2019a). Bayesian graph convolutional neural networks for semi-supervised classification. In *The AAAI Conference on Artificial Intelligence, 33*.

Zhang, Z., & Lee, W.S. (2019). Deep graphical feature learning for the feature matching problem. In *IEEE International Conference on Computer Vision*.

Zhang, Z., Xiang, Y., Wu, L., Xue, B., & Nehorai, A. (2019b). Kergm: Kernelized graph matching. In *International Conference on Neural Information Processing Systems*.

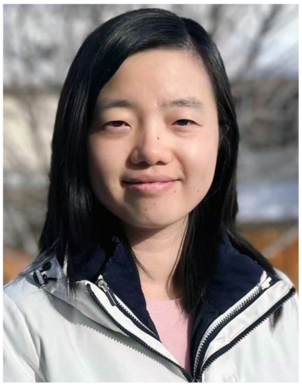
Zhao, R., Oyang, W., & Wang, X. (2016). Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(2), 356–370.

Zhao, R., Wang, K., Su, H., & Ji, Q. (2019a). Bayesian graph convolution LSTM for skeleton based action recognition. In *IEEE International Conference on Computer Vision*.

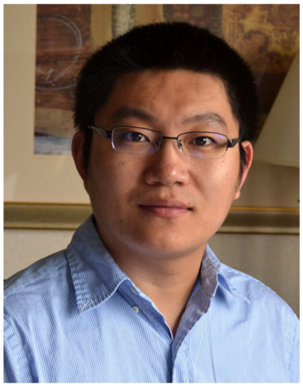
Zhao, Y., Shen, X., Jin, Z., Lu, H., & Hua, X. S. (2019b). Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Peng Gao is a Postdoc associate at the University of Maryland, College Park. He obtained his Ph.D. degree in the computer science department at the Colorado School of Mines. He received his master's degree in automation in Southeast university. His research interests include collaborative perception, deep graph learning, lifelong autonomy, and connected autonomous driving.



Qingzhao Zhu obtained her master degree in the Department of Computer Science at the Colorado School of Mines. Her research interests include collaborative perception in connected vehicles, robot learning and adaptation, and multi-robot systems. She received a master's degree from Peking University.



Hao Zhang is an Associate Professor in the College of Information and Computer Sciences at UMass Amherst, where he directs the Human-Centered Robotics Laboratory. Previously, he was an Assistant Professor (2014–2020) and Associate Professor (2020–2022) of Computer Science at the Colorado School of Mines. Professor Zhang received his PhD from the University of Tennessee Knoxville in 2014, an MS from the Chinese Academy of Sciences in 2009, and a BS from the University of Science and Technology of China (USTC) in 2006.