Bayesian Deep Graph Matching for Correspondence Identification in Collaborative Perception

Peng Gao and Hao Zhang
Human-Centered Robotics Lab, Colorado School of Mines
gaopeng@mines.edu, hzhang@mines.edu

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 in collaborative perception that is caused by occlusion and limited fields of view of the agents, 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 upon graph representations that are constructed from the agents’ observations. We introduce a novel deep graph matching network under the Bayesian framework to explicitly quantify uncertainty in the identified correspondences. In addition, we design a novel loss function that explicitly reduces correspondence uncertainty and perceptual non-covisibility during learning. We evaluate our approach in the robotics applications of collaborative assembly and multi-robot coordination using high-fidelity simulations and physical robots. Experiments have shown that, through addressing both uncertainty and non-covisibility, our approach achieves the state-of-the-art performance of correspondence identification.

I. INTRODUCTION

Collaborative robotics, including multi-robot systems [4] [38] and human-robot collaboration [33, 37], 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 [18] [20], multi-robot search and rescue [1] [45], and connected autonomous driving [19] [49]. 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 [14] [17] [43]. For example, as illustrated by Figure 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 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 [57] [59] and learning-free graph matching [5] [6]. 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) [41] [52], the methods for object reidentification identified the same objects in different frames and from different perspectives [22] [25] [36] [42] [46]. By encoding spatial relationships of the objects using graph neural networks (GNN) [11] [44], deep graph matching was designed to learn graph similarities [48] [55] and graph representations [12] [21] 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 [15]. Without the capability of explicitly quantify and addressing uncertainties during learning, deep graph matching is not robust to noisy observations [24]. 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 (Figure 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.

The remainder of the paper is organized as follows. In Section II, we review existing techniques for correspondence identification. In Section III, we introduce the proposed Bayesian deep graph matching approach. In Section IV, we present and discuss our experimental results in collaborative assembly and multi-robot cooperation applications. Finally, we conclude the paper in Section V.

II. RELATED WORK

A. 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 [9], global [57], or semantic features [59]. The second category of methods use the spatial similarity among objects using, e.g., distances between the objects in pairwise graph matching [6, 29], angular relationships of objects in hypergraph matching [44, 42], spatial relationships built by four or more objects in clique matching [55], and a combination of multiple spatial relationships [5]. The third category of methods recognize object correspondences by enforcing the circle-consistent constraints in multiple views [10], e.g., based on convex relaxation [3], spectral relaxation [32] and graph clustering [50].

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 [17], which addresses the observation uncertainty by adding regularization terms into the graph matching formulation. However, this method can not address the uncertainty in the graph matching model, and is not able to quantify the correspondence uncertainty caused by the perception uncertainty.

B. 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 [48, 55]. The identified correspondence can be improved by designing representative graphs [21] or by removing the correspondences violating neighborhood consensus [12]. The accuracy of deep graph matching can be improved by incorporating combinatorial solvers [39], and the efficiency can be improved by decomposing large graphs into small parts [30]. 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.

C. 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 [23], variance [16] and entropy [31].

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 [13, 27]. The shortcoming of the non-Bayesian
Bayesian-based techniques focus on modeling the distribution of network parameters for uncertainty quantification, including Markov Chain Monte Carlo (MCMC) and Bayes by backprop (BBB) and Monte Carlo Dropout (MC dropout). The Bayesian-based techniques are widely used in various applications, such as using Bayesian GNN with Dirichlet prior and Gaussian prior for node classification and graph classification.

The initial embedding is defined as $h_0^i = a_i$. In collaborative perception, observations acquired by a pair of robots are represented as two graphs $G(V, A, E)$ and $G'(V', A', E')$, respectively. We calculate their respective embedding vectors $H$ and $H'$ using Eq. (1). Then, the visual-spatial similarity of $G$ and $G'$ can be computed as follows:

$$S = HH'^T = \Psi(A, E)\Psi^T(A', E')$$

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

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

$$\arg\max_Y S^TY \quad \text{s.t.} \quad Y_{1:n'\times1} \leq 1_{n\times1}, Y^T1_n \leq 1_{n'\times1}$$ (4)

where $Y = \{y_{ii'}\}$ denotes the correspondence matrix, with $y_{ii'} = 1$ meaning that the $i$-th object in $G$ corresponds to the $i'$-th object in $G'$, and $1$ is a vector with all ones. Eq. (4) aims to maximize the overall similarity of objects’ embedding given the correspondence matrix $Y$. The constraints are used to guarantee one-to-one correspondences by enforcing each row and column in $Y$ 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, which is efficient and strict with one-to-one correspondence constraint.

**B. 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 $W$ in a distribution form instead of taking fixed values. Given a set of $N$ training
instances $\mathcal{X} = \{G_i^*, G_j^*\}^N$ with ground truth $\mathcal{Y} = \{Y_i^*\}^N$, $W$ is computed as:
\[
p(W|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, W)p(W)}{p(\mathcal{Y}|\mathcal{X})}
\]
where $p(W|\mathcal{X}, \mathcal{Y})$ is the posterior distribution of $W$ estimated from its prior distribution $p(W)$. Given $p(W|\mathcal{X}, \mathcal{Y})$, the inference process is defined as follows:
\[
p(\mathcal{Y}|G, G', \mathcal{X}, \mathcal{Y}) = \int_{W\in\Omega} p(\mathcal{Y}|S)p(S|G, G', W)p(W|\mathcal{X}, \mathcal{Y})dW
\]
Under our framework of Bayesian learning, $p(\mathcal{Y}|G, G', \mathcal{X}, \mathcal{Y})$ represents the correspondence matrix $Y$ in a distribution form, rather than taking fixed values through marginalizing over the posterior $p(W|\mathcal{X}, \mathcal{Y})$. $p(\mathcal{Y}|S)$ denotes the probability of $Y$ given $S$, and $p(S|G, G', W)$ denotes the probability of $S$ given the pair of graphs $G, G'$ as input and the model parameter $W$.

Directly computing the integral in Eq. (5) requires to exploit over all the parameter space $\Omega$, which is intractable for the gradient descent-based inference. In order to address this challenge, we adopt the dropout variance inference [16] to obtain the approximated posterior distribution $q(W)$ instead of $p(W|\mathcal{X}, \mathcal{Y})$ by minimizing the Kullback-Leibler divergence:
\[
\min_\theta KL(q_\theta(W)||p(W|\mathcal{X}, \mathcal{Y})) = \min_\theta \int_{W\in\Omega} q_\theta(W) \log \frac{q_\theta(W)}{p(W|\mathcal{X}, \mathcal{Y})}
\]
where $\theta = \{M_1, M_2, \ldots, M_N\}$ denotes the variational parameter with $M_i$ 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 $W_i$ from $q_\theta(W)$ using dropout as follows:
\[
W_i = M_i \cdot \text{diag}([z_{i,j}]_{j=1}^{K_{i-1}})
\]
\[z_{i,j} \sim \text{Bernoulli}(p_i), i = 1, 2, \ldots, L, j = 1, 2, \ldots, K_{i-1}\]

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 $W$. That is, the distribution of correspondence is inferred by:
\[
p(\mathcal{Y}|G, G', \mathcal{X}, \mathcal{Y}) \approx \frac{1}{T} \sum_{i=1}^{T} p(\mathcal{Y}|S)p(S|G, G', W^{(t)}), W^{(t)} \sim q(W)
\]
where $T$ is the number of sampling. We define the final correspondence as the expectation of the correspondence samples sampled from Eq. (8), which is denoted as $E(p(\mathcal{Y}))$, where $E$ denotes the expectation function. The uncertainty of each correspondence is defined as follows:
\[
\mathbb{H}(E(p(\mathcal{Y}))) = -E(p(\mathcal{Y})) \ast \log(E(p(\mathcal{Y})))
\]
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 [8].

The loss function for our Bayesian deep graph matching approach is defined as follows:
\[
\mathcal{L}_{\text{coid}} = -\log \left( \frac{1}{nn'} || \mathbf{S} \circ Y^* \circ E(Y) ||_1 \right)
\]
where $\circ$ represents the element-wise product, $n$ and $n'$ are the number of objects in graph $G$ and $G'$ respectively, and $Y^*$ denotes the ground truth of the correspondence matrix, with $Y^*_{i,i'} = 1$ denoting the ground truth of correspondence between the $i$-th object in graph $G$ and the $i'$-th object in graph $G'$. 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 [16], minimizing the negative-log loss function $\mathcal{L}_{\text{coid}}$ 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.

C. 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:
\[
\mathcal{L}_{\text{non}} = -\log \left( \frac{1}{nn'} \exp (-\mathbf{S} \circ \mathbf{N} \circ E(Y)) ||_1 \right)
\]
where $\mathbf{N} \in \mathcal{R}^{n \times n'}$ denotes an indicator matrix that includes the indices of non-covisible objects in $Y$, with $\mathbf{N}_{i,i'} = 1$ indicating that the correspondence $Y^*_{i,i'}$ is constructed by non-covisible objects. For example, if the $i$-th object in graph $G$ or the $i'$-th object in graph $G'$ 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 $\mathbf{S} \circ \mathbf{N} \circ E(Y)$. 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 $\mathcal{L}_{\text{coid}}$ to improve the robustness of deep graph matching, which is defined as:
\[
\mathcal{L}_{\text{unc}} = -\log \left( \frac{1}{nn'} \exp (-\mathbb{H}(E(Y))) ||_1 \right)
\]
where $\mathbb{H}(E(Y))$ is our quantified uncertainty in the identified correspondences.

Our final loss function is represented as $\mathcal{L} = \mathcal{L}_{\text{coid}} + \mathcal{L}_{\text{non}} + \mathcal{L}_{\text{unc}}$. Minimizing this loss function during training is equivalent to maximizing the similarity of correct correspondences and minimizing the similarity of non-covisible
objects and matching uncertainty. During execution, given the quantified uncertainty in the identified correspondence, we further improve the correspondences results by defining a threshold $\lambda$, in order to remove the correspondences with high uncertainty values \cite{17}. Specifically, if $\mathbb{E}(p(Y)_{i,j}) \geq \lambda$, the correspondence $Y_{i,j}$ is removed.

IV. 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.

A. 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 Figure 2(a), Real-world furniture assembly tasks (RFAT) as shown in Figure 2(b) and Simulated multi-robot coordination (SMRC) as shown in Figure 2(c).

We construct each observation as a graph with node attributes generated from appearance features \cite{17}. 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 $\Psi$, 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 $\varphi$, 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 $\mathcal{L} = \mathcal{L}_{\text{non}} + \mathcal{L}_{\text{unc}} + \mathcal{L}_{\text{cov}}$ as the loss function. We also implement two baseline methods, using $\mathcal{L}_{\text{cov}} + \mathcal{L}_{\text{non}}$ that addresses only non-covisibility, and $\mathcal{L}_{\text{cov}} + \mathcal{L}_{\text{unc}}$ 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) \cite{5}, which integrates multiple different attributes in a learning-free way to identify correspondences.
- Regularized graph matching (RGM) \cite{17}, which addresses perception uncertainty and non-covisible objects in a learning-free way to identify correspondences.
- Graph convolutional network-based graph matching (GCN-GM) \cite{11}, which identifies correspondences by only optimizing the loss of overall similarity between two observations.
- Deep graph matching consensus (DGMC) \cite{12}, 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 \cite{6,17}, 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>SFAT Recall</th>
<th>SFAT Precision</th>
<th>RFAT Recall</th>
<th>RFAT Precision</th>
<th>SMRC Recall</th>
<th>SMRC Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOGM \cite{5}</td>
<td>0.4385</td>
<td>0.2332</td>
<td>0.2298</td>
<td>0.2467</td>
<td>0.7184</td>
<td>0.7136</td>
</tr>
<tr>
<td>RGM \cite{17}</td>
<td>0.4434</td>
<td>0.2841</td>
<td>0.2871</td>
<td>0.3012</td>
<td>0.7878</td>
<td>0.7735</td>
</tr>
<tr>
<td>GCN-GM \cite{11}</td>
<td>0.9078</td>
<td>0.5398</td>
<td>0.7580</td>
<td>0.8916</td>
<td>0.9321</td>
<td>0.8481</td>
</tr>
<tr>
<td>DGM \cite{12}</td>
<td>0.9105</td>
<td>0.5841</td>
<td>0.9053</td>
<td>0.8971</td>
<td>0.9388</td>
<td>0.9037</td>
</tr>
<tr>
<td>$L_{\text{cov}} + L_{\text{non}}$</td>
<td>0.9123</td>
<td>0.7526</td>
<td>0.9960</td>
<td>0.9036</td>
<td>0.9477</td>
<td>0.9719</td>
</tr>
<tr>
<td>$L_{\text{cov}} + L_{\text{unc}}$</td>
<td>0.9053</td>
<td>0.7011</td>
<td>0.9037</td>
<td>0.9038</td>
<td>\textbf{0.9529}</td>
<td>0.9611</td>
</tr>
<tr>
<td>Ours</td>
<td>\textbf{0.9216}</td>
<td>\textbf{0.7026}</td>
<td>0.9920</td>
<td>\textbf{0.9498}</td>
<td>0.9503</td>
<td>\textbf{0.9683}</td>
</tr>
</tbody>
</table>

B. 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.
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.]

Human view
Robot view
Robotic assistance
Completed task

Scenario on human-robot collaborative assembling task

Step 1  Step 2  Step 3  Step 4  Step 5  Step 6

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.

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 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. 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 Figure 3(c). We can see that our approach can accurately identify correspondences. Compared with GCN-GM and DGMC as shown in Figure 3(a) and Figure 3(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 I. We observe that our baseline methods $L_{coid} + L_{non}$ and $L_{coid} + L_{unc}$ 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.

C. Results in Real-world Furniture Assembly Scenarios

Our approach is further evaluated on RFAT, in which a human and a robot collaboratively assemble 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 [51]. 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 Figure 3[f]. 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 I. We can see that our baseline methods $L_{\text{covid}} + L_{\text{non}}$ and $L_{\text{covid}} + L_{\text{unc}}$ 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.

D. 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 Figure 3[d]. We observe that our approach can correctly identify the correspondences. The results of GCN-GM and DGMC are shown in Figure 3[g] and Figure 3[h] separately. It is observed that the objects far away from the camera are identified incorrectly due to 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 II. 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Before threshold</th>
<th>After threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFAT</td>
<td>0.7009</td>
<td>0.8303</td>
</tr>
<tr>
<td>RFAT</td>
<td>0.9695</td>
<td>0.9724</td>
</tr>
<tr>
<td>SMRC</td>
<td>0.9456</td>
<td>0.9686</td>
</tr>
</tbody>
</table>

E. Discussion

We further evaluate various characteristics of our approach, including the importance of uncertainty quantification in correspondence identification, the performance of our approach using different uncertainties, and hyperparameter analysis.

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 $\lambda$, in order to remove the correspondences with high uncertainty values. As shown in Table III, 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SFAT</th>
<th>RFAT</th>
<th>SMRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epistemic [8]</td>
<td>0.7009</td>
<td>0.9722</td>
<td>0.9456</td>
</tr>
<tr>
<td>Aleatoric [8]</td>
<td>0.8303</td>
<td>0.9695</td>
<td>0.9676</td>
</tr>
<tr>
<td>Shannon Entropy [8]</td>
<td>0.8143</td>
<td>0.9724</td>
<td>0.9688</td>
</tr>
</tbody>
</table>

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) [8]. Shannon entropy represents the total uncertainty, as defined in Eq. (10). Given the F1 scores reported in Table III, we can see that using aleatoric 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 uncertainty achieves the best performance in SFAT, which indicates due to the representation of both model and data uncertainty.

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 uncertainty achieves the best performance in SFAT, which indicates due to the representation of both model and data uncertainty.

3) Hyperparameter Analysis: We use the hyperparameter $\lambda$ 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 $\lambda$ based on the F1 score. As shown in Figure 6(a), 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 Figure 6(b), 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 $[80, 90]$.

V. 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. Extensive experiments are conducted to evaluate our method in collaborative furniture assembly and multi-robot coordination 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.

VI. ACKNOWLEDGMENTS

This work was partially supported by NSF CAREER Award IIS-1942056. The authors also thank the anonymous reviewers and the area chair for their feedback.
REFERENCES


[55] Zhen Zhang and Wee Sun Lee. Deep graphical feature


