Robotics: Science and Systems 2021
Held Virtually, July 12–16, 2021

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, 7, 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.

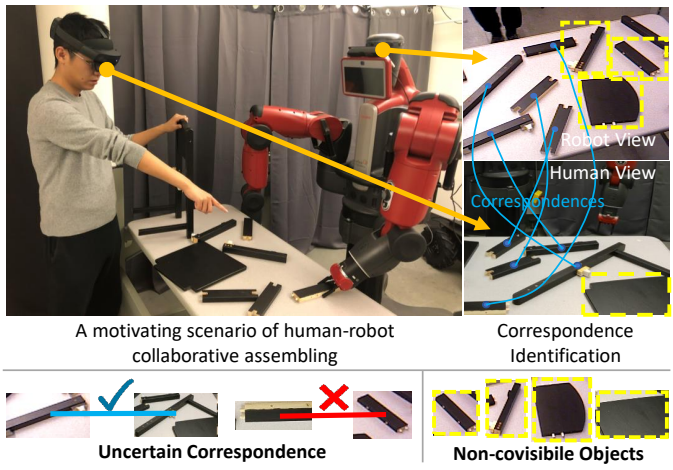


Fig. 1. This example motivates correspondence identification in collaborative perception in the application of human-robot collaborative assembly. When a collaborative robot assists a human worker who wears an augmented reality (AR) headset to assemble a chair, they must identify the correspondence of 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.

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, 47, 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 quantifying 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 [34, 42], spatial relationships built by four or more objects in clique matching [35], 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], vacuity and dissonance uncertainty from subjective logic perspective [12], 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

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) [26], Bayes by backprop (BBB) [2] and Monte Carlo Dropout (MC dropout) [16]. The Bayesian-based techniques are widely used in various applications, such as using Bayesian GNN with Dirichlet prior [31, 55] and Gaussian prior [40] for node classification [54], edge prediction [53] and graph classification [58].

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.

III. 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.

A. 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 positions of the objects detected in the observation, where $\mathbf{v}_i \in \mathcal{R}^{d_v}$ is the position 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. 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 $\mathbf{v}_j$.

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 $\mathbf{H} = \Psi(\mathbf{A}, \mathbf{E})$, 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. The initial embedding is defined as $\mathbf{h}_i^0 = \mathbf{a}_i$.

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 [12], 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:

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

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. Eq. (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 [56, 12] that 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 $\mathbf{W}$ 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:

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

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 $p(\mathbf{W}|\mathcal{X}, \mathcal{Y})$. $p(\mathbf{Y}|\mathbf{S})$ denotes the probability of $\mathbf{Y}$ given $\mathbf{S}$, and $p(\mathbf{S}|\mathcal{G}, \mathcal{G}', \mathbf{W})$ denotes the probability of $\mathbf{S}$ 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 exploit 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 [16] 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:

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

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:

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

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 [8].

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 [16], 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.

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:

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

where $\mathbf{N} \in \mathcal{R}^{n\times n'}$ 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:

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

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

(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.

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 [17]. Specifically, if $\mathbb{H}(\mathbb{E}(p(\mathbf{Y})_{i,i'})) \geq \lambda$, the correspondence $\mathbf{Y}_{i,i'}$ 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 [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 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 $\mathcal{L}_{coid} + \mathcal{L}_{non}$ that addresses only non-covisibility, and $\mathcal{L}_{coid} + \mathcal{L}_{unc}$ 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**) [5], which integrates multiple different attributes in a learning-free way to identify correspondences.
- Regularized graph matching (**RGM**) [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**) [11], which identifies correspondences by only optimizing the loss of overall similarity between two observations.
- Deep graph matching consensus (**DGMC**) [12], which uses the similarity of embedding vectors obtained by graph neural networks for correspondence identification while checking the neighborhood consensus of identified correspondences.

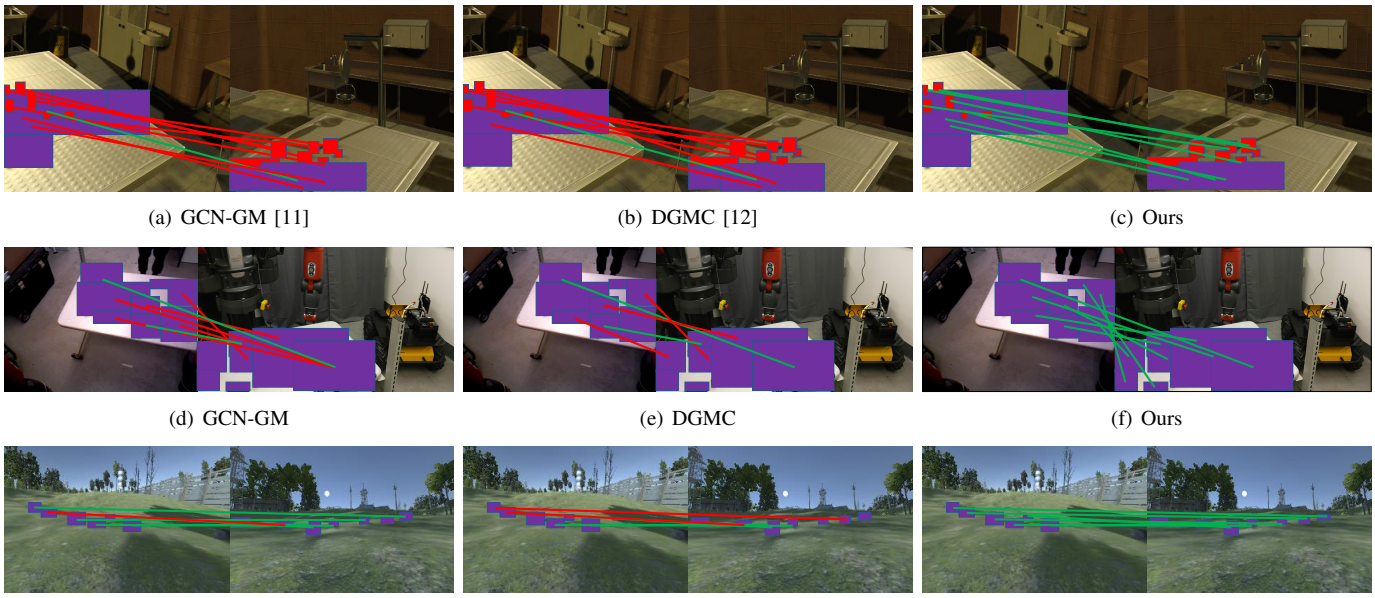
Following a standard experimental setup [6, 17], precision and recall are adopted to evaluate our approach. Given the identified correspondences, precision is defined as the ratio of correct correspondences over all the identified correspondences. Recall is defined as the ratio of identified correspondences 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 I
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 [5]	0.4385	0.2332	0.2298	0.2467	0.7184	0.7136
RGM [17]	0.4434	0.2841	0.2871	0.3012	0.7878	0.7735
GCN-GM [11]	0.9078	0.5398	0.7580	0.8916	0.9321	0.8481
DGMC [12]	0.9105	0.5441	0.9933	0.8971	0.9388	0.9037
$\mathcal{L}_{coid} + \mathcal{L}_{non}$	0.9122	0.5526	**0.9960**	0.9036	0.9477	0.9319
$\mathcal{L}_{coid} + \mathcal{L}_{unc}$	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**

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.



(a) GCN-GM [11] (b) DGMC [12] (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.]

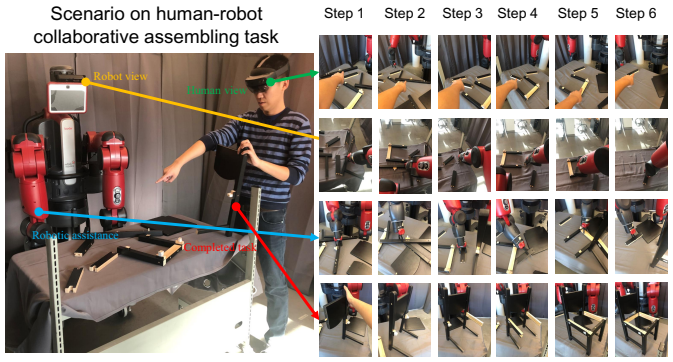


Fig. 4. Illustrations of several steps in the scenario of robot-assisted furniture assembly. The Baxter robot assists a human collaborator who wears an AR headset to collaboratively assemble an IKEA chair.

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 [28]. 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 $\mathcal{L}_{coid} + \mathcal{L}_{non}$ 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.

C. 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 [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 $\mathcal{L}_{coid} + \mathcal{L}_{non}$ 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.

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(i). 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 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 I. 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 II
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**

E. Discussion

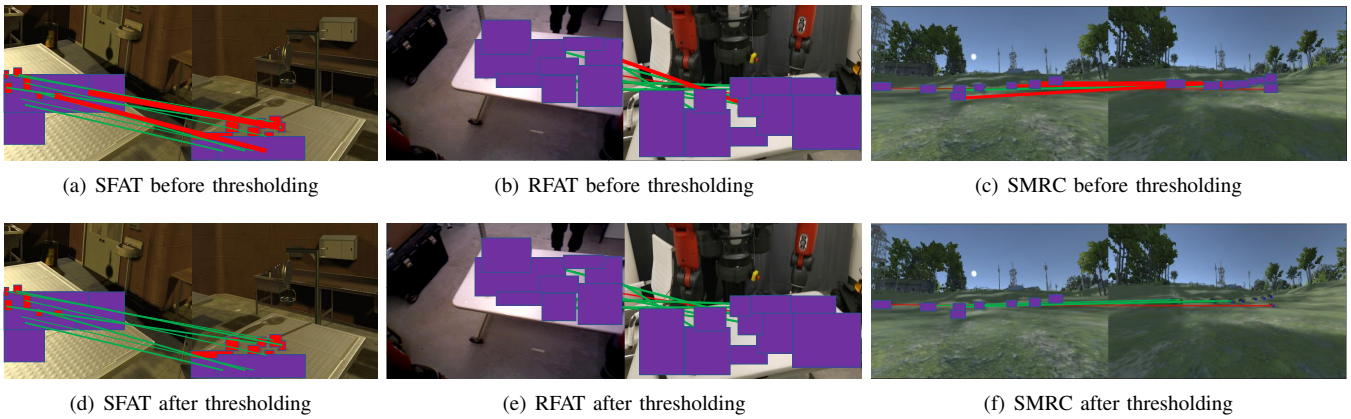
We further evaluate various characteristics of our approach, including the importance of uncertainty quantification in correspondence identification, the performance of our approach using different uncertainties, and hyperparameter analysis.

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 λ, in order to remove the correspondences with high uncertainty values. As shown in Table II, 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 III
QUANTITATIVE ANALYSIS ON THE PERFORMANCE OF OUR APPROACH USING DIFFERENT TYPES OF UNCERTAINTY. THE METRIC REPORTED IS THE F1-SCORE OVER SFAT, RFAT AND SMRC.

Methods	SFAT	RFAT	SMRC
Epistemic [8]	0.7009	0.9722	0.9456
Aleatoric [8]	**0.8303**	0.9695	0.9676
Shannon Entropy [8]	0.8143	**0.9724**	**0.9688**

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

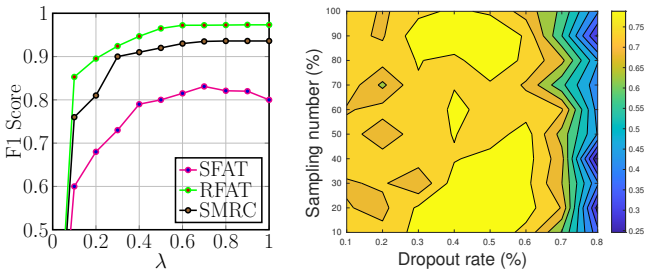


(a) SFAT before thresholding (b) RFAT before thresholding (c) SMRC before thresholding

(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.]

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 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.



(a) Threshold effects (b) Dropout rate and sampling number

Fig. 6. Hyperparameter analysis based on the metric of F1 scores.

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 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

[1] José J Acevedo, Joao Messias, Jesús Capitán, Rodrigo Ventura, Luis Merino, and Pedro U Lima. A Dynamic Weighted Area Assignment Based on a Particle Filter for Active Cooperative Perception. *IEEE Robotics and Automation Letters*, 5(2):736–743, 2020.

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[4] Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm robotics: A review from the swarm engineering perspective. *Swarm Intelligence*, 7 (1):1–41, 2013.

[5] Hyung Jin Chang, Tobias Fischer, Maxime Petit, Martina Zambelli, and Yiannis Demiris. Learning kinematic structure correspondences using multi-order similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2920–2934, 2017.

[6] Minsu Cho, Jungmin Lee, and Kyoung Mu Lee. Reweighted random walks for graph matching. In *European Conference on Computer Vision*, 2010.

[7] Soon-Jo Chung, Aditya Avinash Paranjape, Philip Dames, Shaojie Shen, and Vijay Kumar. A survey on aerial swarm robotics. *IEEE Transactions on Robotics*, 34(4):837–855, 2018.

[8] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Uncertainty decomposition in bayesian neural networks with latent variables. *arXiv preprint*, 2017.

[9] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*, 2014.

[10] Kaveh Fathian, Kasra Khosoussi, Yulun Tian, Parker Lusk, and Jonathan P How. CLEAR: A consistent lifting, rmbedding, and alignment rectification algorithm for multi-agent data association. *IEEE Transactions on Robotics*, 36(6):1686–1703, 2020.

[11] Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. SplineCNN: Fast geometric deep learning with continuous b-spline kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Matthias Fey, Jan E Lenssen, Christopher Morris, Jonathan Masci, and Nils M Kriege. Deep Graph Matching Consensus. In *International Conference on Learning Representations*, 2019.

[13] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv*, 2019.

[14] Kristoffer M Frey, Ted J Steiner, and Jonathan P How. Efficient constellation-based map-merging for semantic SLAM. In *IEEE International Conference on Robotics and Automation*, 2019.

[15] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *International Conference on Neural Information Processing Systems*, 2015.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

[17] Peng Gao, Rui Guo, Hongsheng Lu, and Hao Zhang. Regularized Graph Matching for Correspondence Identification under Uncertainty in Collaborative Perception. *Robotics: Science and Systems*, 2020.

[18] Peng Gao, Brian Reily, Savannah Paul, and Hao Zhang. Visual reference of ambiguous objects for augmented reality-powered human-robot communication in a shared workspace. In *International Conference on Human-Computer Interaction*, 2020.

[19] Rui Guo, Hongsheng Lu, Peng Gao, Ziling Zhang, and Hao Zhang. Collaborative localization for occluded objects in connected vehicular platform. In *IEEE 90th Vehicular Technology Conference*, 2019.

[20] Antti Hietanen, Roel Pieters, Minna Lanz, Jyrki Latokartano, and Joni-Kristian Kämäräinen. AR-based interaction for human-robot collaborative manufacturing. *Robotics and Computer-Integrated Manufacturing*, 63: 101891, 2020.

[21] Bo Jiang, Pengfei Sun, Jin Tang, and Bin Luo. Glmnet: Graph learning-matching networks for feature matching. *arXiv*, 2019.

[22] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI Conference on Artificial Intelligence*, 2020.

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *International Conference on Neural Information Processing Systems*, 2017.

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] Amena Khatun, Simon Denman, Sridha Sridharan, and Clinton Fookes. Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification. In *IEEE Winter Conference on Applications of Computer Vision*, 2020.

[26] Matthew A Kupinski, John W Hoppin, Eric Clarkson, and Harrison H Barrett. 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, 2003.

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty

estimation using deep ensembles. In *International Conference on Neural Information Processing Systems*, 2017.

[28] Youngwoon Lee, Edward S Hu, Zhengyu Yang, Alex Yin, and Joseph J Lim. IKEA furniture assembly environment for long-horizon complex manipulation tasks. *arXiv*, 2019.

[29] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *IEEE International Conference on Computer Vision*, 2005.

[30] Zhaoyu Lou, Jiaxuan You, Chengtao Wen, Arquimedes Canedo, Jure Leskovec, et al. Neural subgraph matching. *arXiv*, 2020.

[31] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *International Conference on Neural Information Processing Systems*, 2018.

[32] Eleonora Maset, Federica Arrigoni, and Andrea Fusiello. Practical and efficient multi-view matching. In *IEEE International Conference on Computer Vision*, 2017.

[33] Sachiko Matsumoto and Laurel D Riek. Fluent coordination in proximate human robot teaming. In *Robotics: Science and Systems workshop*, 2019.

[34] Quynh Nguyen, Antoine Gautier, and Matthias Hein. A flexible tensor block coordinate ascent scheme for hypergraph matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[35] Wei-Zhi Nie, An-An Liu, Zan Gao, and Yu-Ting Su. Clique-graph matching by preserving global & local structure. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[36] Rodolfo Quispe and Helio Pedrini. Improved person re-identification based on saliency and semantic parsing with deep neural network models. *Image and Vision Computing*, 92:103809, 2019.

[37] Brian Reily, Fei Han, Lynne E Parker, and Hao Zhang. Skeleton-based bio-inspired human activity prediction for real-time human–robot interaction. *Autonomous Robots*, 42(6):1281–1298, 2018.

[38] Brian Reily, Christopher Reardon, and Hao Zhang. Representing multi-robot structure through multi-modal graph embedding for the selection of robot teams. *arXiv*, 2020.

[39] Michal Rolínek, Paul Swoboda, Dominik Zietlow, Anselm Paulus, Vít Musil, and Georg Martius. Deep graph matching via blackbox differentiation of combinatorial solvers. *European Conference on Computer Vision*, 2020.

[40] Seongok Ryu, Yongchan Kwon, and Woo Youn Kim. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chemical Science*, 10(36):8438–8446, 2019.

[41] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, 2016.

[42] Yumin Suh, Kamil Adamczewski, and Kyoung Mu Lee. Subgraph matching using compactness prior for robust feature correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[43] Yulun Tian, Katherine Liu, Kyel Ok, Loc Tran, Danette Allen, Nicholas Roy, and Jonathan P How. Search and rescue under the forest canopy using multiple UAVs. *The International Journal of Robotics Research*, 2019.

[44] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[45] Alberto Viseras, Zhe Xu, and Luis Merino. Distributed multi-robot cooperation for information gathering under communication constraints. In *IEEE International Conference on Robotics and Automation*, 2018.

[46] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[47] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[48] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *IEEE International Conference on Computer Vision*, 2019.

[49] ShangGuan Wei, Du Yu, Chai Lin Guo, Liu Dan, and Wang Wei Shu. Survey of connected automated vehicle perception mode: from autonomy to interaction. *Intelligent Transport Systems*, 13(3):495–505, 2018.

[50] Junchi Yan, Zhe Ren, Hongyuan Zha, and Stephen Chu. A constrained clustering based approach for matching a collection of feature sets. In *International Conference on Pattern Recognition*, 2016.

[51] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[52] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[53] Baichuan Zhang, Sutanay Choudhury, Mohammad Al Hasan, Xia Ning, Khushbu Agarwal, Sumit Purohit, and Paola Pesntez Cabrera. Trust from the past: Bayesian personalized ranking based link prediction in knowledge graphs. *arXiv*, 2016.

[54] Yingxue Zhang, Soumyasundar Pal, Mark Coates, and Deniz Ustebay. Bayesian graph convolutional neural networks for semi-supervised classification. In *The AAAI Conference on Artificial Intelligence*, volume 33, 2019.

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

learning for the feature matching problem. In *IEEE International Conference on Computer Vision*, 2019.

[56] Zhen Zhang, Yijian Xiang, Lingfei Wu, Bing Xue, and Arye Nehorai. Kergm: Kernelized graph matching. In *International Conference on Neural Information Processing Systems*, 2019.

[57] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):356–370, 2016.

[58] Rui Zhao, Kang Wang, Hui Su, and Qiang Ji. Bayesian graph convolution LSTM for skeleton based action recognition. In *IEEE International Conference on Computer Vision*, 2019.

[59] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.