Infinite Latent Conditional Random Fields for Modeling Environments through Humans

Yun Jiang and Ashutosh Saxena.
Department of Computer Science, Cornell University, USA.
{yunjiang,asaxena}@cs.cornell.edu

***Abstract*—Humans cast a substantial influence on their environments by interacting with it. Therefore, even though an environment may physically contain only objects, it cannot be modeled well without considering humans. In this paper, we model environments not only through objects, but also through latent human poses and human-object interactions. However, the number of potential human poses is large and unknown, and the human-object interactions vary not only in type but also in which human pose relates to each object.**

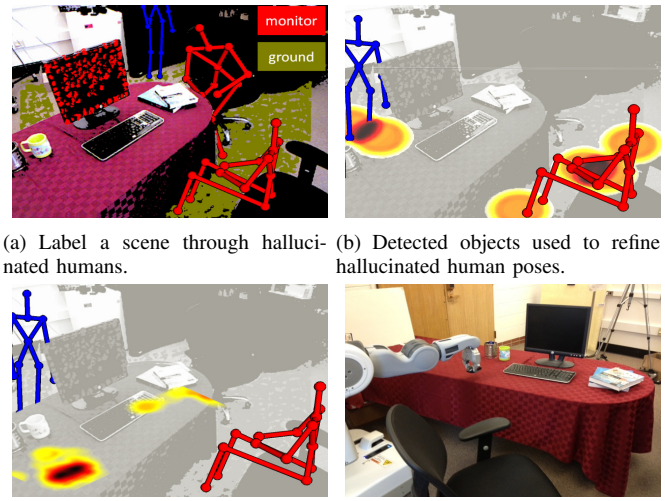
In order to handle such properties, we present Infinite Latent Conditional Random Fields (ILCRFs) that model a scene as a mixture of CRFs generated from Dirichlet processes. Each CRF represents one possible explanation of the scene. In addition to visible object nodes and edges, it generatively models the distribution of different CRF structures over the latent human nodes and corresponding edges. We apply the model to the challenging application of robotic scene arrangement. In extensive experiments, we show that our model significantly outperforms the state-of-the-art results. We further use our algorithm on a robot for placing objects in a new scene.

I. Introduction

That the human environments and the objects in it are designed for human usage, is so deeply ingrained in us that when we think about a human environment, we think it through the interplay between these elements. When a robot is to perform tasks in an environment such as scene labeling and scene arrangement, it also needs to model the environment *through* humans. For example, when a robot is asked to find the monitor in a room (Fig. 1), if it understands how human interact with a monitor, it would scan the table top more carefully than other places. Now, if it is asked to place a mouse, considering how human interact with the mouse and monitor, it would put it at accessible places such as the right front of the monitor.

A human environment is constructed under two types of relations: *object-object* and *human-object relations*. When only considering object-object relations, Conditional random fields (CRFs) are a natural choice, as each object can be modeled as a node in a Markov network and the edges in the graph can reflect the object-object relations. CRFs and their variants have thus been applied to many scene modeling tasks (e.g., [28, 1, 25], see more related work in § IV-C).

Modeling possible human poses and human-object interactions (or *object affordances*) is not trivial because of several reasons. First, humans are not always observable, but we still want to model them as latent factors for making the scene as it is. Second, there can be any number of possible humans in a scene—e.g., some sitting on the couch/chair, some standing by the shelf/table; Third, there can be various types of human-object interactions in a scene, such as watching TV in distance, eating from dishes, or working on a laptop, etc; Fourth, an object can be used by different human poses, such as a book on the table can be accessed by either a sitting pose on the couch or a standing pose nearby; Last, there can be multiple possible usage scenarios in a scene (e.g., see Figure 2-middle row). Therefore, we need models that can incorporate latent factors, latent structures, as well as different alternative possibilities.



(a) Label a scene through hallucinated humans.

(b) Detected objects used to refine hallucinated human poses.

(c) Infer locations for placing a mouse (shown in heat map).

(d) Execute the placing task.

Fig. 1: Robotic experiment: Given a RGB-D scene, our robot first labels the segments in the point cloud using hallucinated humans (a). Then, when asked to arrange a mouse in the scene, it infers possible human poses (b), as well as proper placements (c) using our ILCRF-based scene arrangement algorithm. Finally, it executes the placing action (d).

In this work, we propose infinite latent conditional random fields (ILCRFs) for modeling the aforementioned properties. Intuitively, it is a mixture of CRFs where each CRF can have two types of nodes: existing nodes (e.g., object nodes, which are given in the graph and we only have to infer the value) and latent nodes (e.g., human nodes, where an unknown number of humans may be hallucinated in the room). The relations between the nodes (object-object edges and human-object edges) could also be of different types. Unlike traditional CRFs, where the structure of the graph is given, the structure

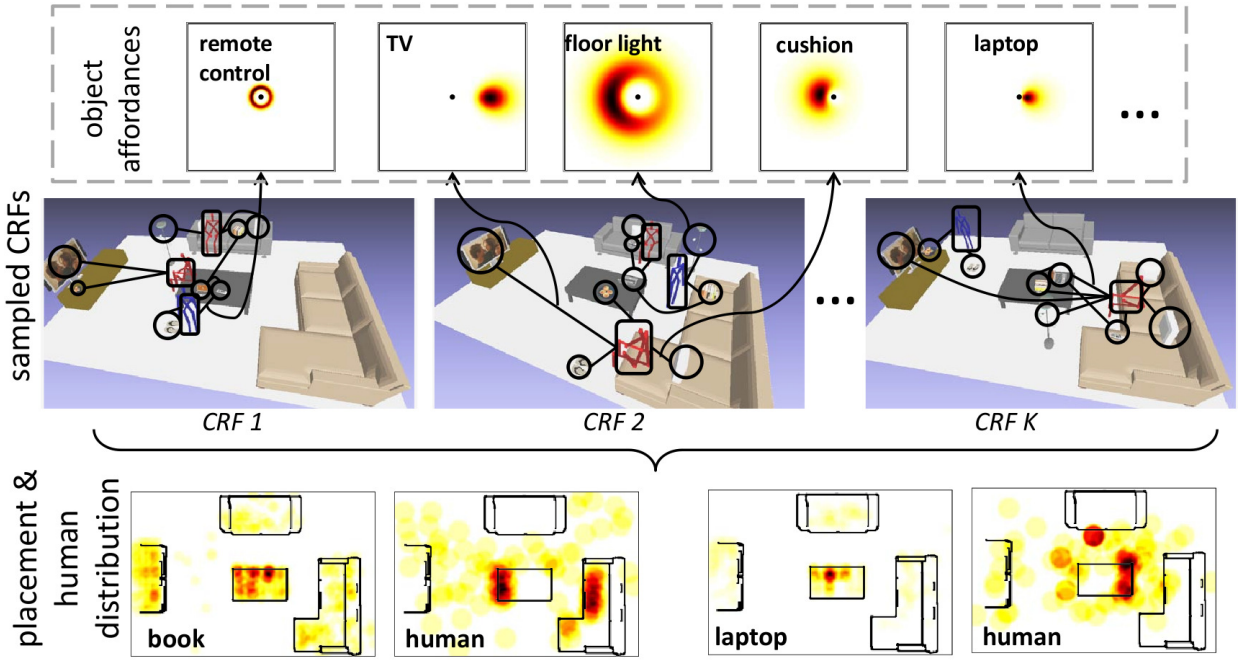


Fig. 2: An example of instantiated ILCRF for scene arrangement. Top row shows learned object affordances in top-view heatmaps (it shows the probability of the object's location, given a human pose in the center facing to the *right*). Middle row shows a total of K CRFs sampled from our ILCRF algorithm—each CRF models the scene differently. Bottom row shows the distribution of the objects and humans (in the top view of the room) computed from the sampled CRFs.

of our ILCRF is sampled from Dirichlet Processes (DPs). DPs are widely used as nonparametric Bayesian priors for mixture models, the resulting DP mixture models can determine the number of components from data, and therefore is also referred as infinite mixture model. ILCRFs are inspired by this, and we call it 'infinite' as it can sidestep the difficulty of finding the correct number of latent nodes as well as latent edge types. Our learning and inference methods are based on Gibbs sampling that samples latent nodes, existing nodes, and edges from their posterior distributions.

We apply our ILCRF to the task of scene arrangement where the objective is to find proper placements (including 3D location and orientation) of given objects in a scene. For this particular application, our ILCRF models each object placement as an existing node, hallucinated human poses as latent nodes and spatial relationships among objects or between objects and humans as edges. We demonstrate in the experiments that this model achieves the state-of-the-art results on both synthetic and real datasets. More importantly, we perform an exhaustive analysis on how our model captures different aspects of human context in scenes, in comparisons with numerous baselines. We further demonstrate that by modeling through latent human context, a robot successfully identified the class of objects in a new room, and placed several objects correctly in it.

In summary, the contributions of this paper are as follows:

- We propose ILCRFs to capture both human-object and object-object relations in a scene where humans are hidden. Previous work [14] only admits modeling human-object relations.
- Compared to classic CRFs, our ILCRFs admit: 1) unknown number of latent variables, 2) unknown number of potential functions, and 3) a mixture of different CRFs. Its flexibility allows us to have minimum restrictions on humans and affordances.

II. Preliminaries: Conditional Random Fields

Definition 1. *$CRF(\mathcal{X}, \mathcal{Y}, E_Y)$ is a conditional random field if that, when conditioned on $\mathcal{X}$, random variables $\mathcal{Y}$ follow the Markov property with respect to the graph E_Y: The absence of an edge between nodes y_i and y_j implies that they are independent given other nodes.* ■

Thus, the likelihood of $\mathcal{Y}$ given $\mathcal{X}$ is given by: $P(\mathcal{Y}|\mathcal{X}) \propto \prod_{c\in\mathcal{C}} \psi_c(X_c, Y_c)$, where $\mathcal{C}$ is all the maximum cliques, and $Y_c \in \mathcal{Y}$ and $X_c \in \mathcal{X}$ are in the same clique c. Figure 3 (top) shows an example of a chained CRF. When the graph structural E_Y is unknown, structural learning is required and the task is to select Y_c for each clique c. While it is known to be hard in general [32], there are heuristic methods and approximate inference for tree-structured CRFs [31, 5].

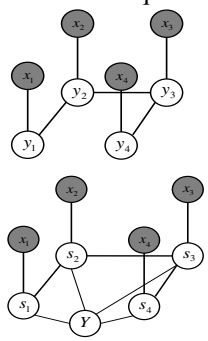


Fig. 3: CRF (top) and hidden CRF (bottom) in graphical representation.

A. Related Work: Variants of CRFs

Variants of Conditional Random Fields (CRFs) ([22]) have emerged as a popular way to model hidden states and have been successfully applied to many vision problems.

There are many models that enrich the structure of labels in CRFs. For example, latent CRFs [26] assume that the overall label Y depends on a sequence of hidden states

$(s_1, s_2, \ldots, s_k)$ (see Fig. 3-bottom). This can be applied to object recognition (an object label is determined by its part labels) [29] and gesture recognition [36]. Further, factorial (or dynamic) CRFs [33] substitute every label with a Markov network structure to allow structured labeling, especially for sequential data (such as labeling object and action simultaneously in video sequences [18, 21]). However, the labels and hidden states are discrete and take only finite number of values. In contemporary work, Bousmalis et al. [4] present a model that shares a name similar to ours, but is quite different. They estimate the correct number of values a latent node can take using Dirichlet processes in a way similar to augmenting hidden Markov models (HMM) to infinite HMM [3]. However, the number of hidden nodes is fixed in their model. In our model, we estimate the number of latent nodes, and even allow the labels to be continuous.

Some works impose a non-parametric Bayesian prior to the network's structure so that it can potentially generate as many nodes as needed. For example, Indian Buffet process [8] assumes the latent nodes and links are generated through Beta processes and the infinite factorial HMM [35] incorporates it to HMM to allow any number of latent variables for each observation. However, they are limited to binary Markov chains and do not consider different types of potential functions either. Thus these models are complementary to ours. Jancsary et al. [12] considers Gaussian CRFs on fixed number of nodes but unknown number of potential functions and proposes a non-parametric method to learn the number as well as parameters of each potential function. Unlike this work, our model can handle unknown number of *nodes* as well as *types of edges*.

Cast in the light of mixture models, mixtures of graphical models have been proposed to overcome the limited representational power that a single graphs often suffers. For example, Anandkumar et al. [2] propose a novel method to estimate a mixture of a finite number of discrete graphs from data. Both Rodriguez et al. [27] and Ickstadt et al. [10] consider a Dirichlet process mixture model over graphs so that the number of different graphical models is determined by the data. However, they are limited to Gaussian graphical models and do not consider latent variables.

III. Infinite Latent Condition Random Fields

In this paper, we propose a type of mixture CRFs—infinite latent conditional random fields (ILCRFs), which can capture the following properties:

1. Unknown number of latent nodes. This is essential for applications of finding hidden causes, such as scene modeling where the number of possible human poses in a scene is unknown and changes across different scenes.

*2. Unknown number of **types** of potential functions.* Potential function measures the relationship between nodes, and therefore, having variety in them can help us model complex relations. For example, in the task of image segmentation, different types of context can be modeled as different edges in a CRF [12]. In this paper, we use them to capture different object affordances.

3. Mixture CRFs. The complexity of real-world data may not always be explained by a single CRF. Therefore having a mixture of CRFs, with each one modeling one particular conditional independency in the data, can increase the expressive power of the model.

4. Ability to place informative priors on the structure of CRFs. This can help producing more plausible CRFs as well as reducing the computational complexity.

We achieve this by imposing Bayesian nonparametric priors—Dirichlet processes (DPs)—to the latent variables, potential functions and graph structures.

A. Background: Dirichlet Process Mixture Model

Dirichlet process is a stochastic process to generate distributions that is used to model clustering effects in the data. It has been widely applied to modeling *unknown* number of components in mixture models, which are often called *infinite* mixture models. (Formal definition can be found in [34].)

Definition 2. *A DP mixture model, $DP(\alpha, B)$, defines the following generative process (also called the stick-breaking process), with a concentration parameter α and a base distribution B:*

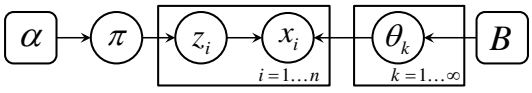
1) *Generate infinite number of mixture components, parameterized by $\Theta = \{\theta_1, \ldots, \theta_\infty\}$, and their mixture weights $\boldsymbol{\pi}$:*

$$\theta_k \sim B,\ b_k \sim Beta(1, \alpha),\ \ \pi_k = b_k \textstyle\prod_{i=1}^{k-1}(1-b_i). \quad (1)$$

2) *Assign the z_i^{th} component to each data point x_i and draw from it:*

$$z_i \sim \boldsymbol{\pi}, \qquad x_i \sim F(\theta_{z_i}). \quad (2)$$

The process can be represented in a plate notation as below:



■

B. ILCRF

ILCRF uses DPs to admit an arbitrary number of latent variables and potential functions to obtain a mixture of latent CRFs. In brief, it generates latent variables and potential functions from two DPs respectively, and each data point builds a link, associated with one potential function, to one latent variable. Different samples thus form different CRFs.

Definition 3. *A ILCRF$(\mathcal{X}, \mathcal{Y}, E_Y, \alpha_h, B_h, \alpha_\psi, B_\psi)$ is a mixture of CRFs, where the edges in $\mathcal{Y}$ are defined in graph E_Y and latent variables $\mathcal{H}$ as well as the edges between $\mathcal{H}$ and $\mathcal{Y}$ are generated through the following process:*

1) *Generate infinite number of latent nodes $\mathcal{H} = \{h_1, h_2, \ldots, h_\infty\}$ and a distribution $\boldsymbol{\pi}_h$ from a DP process $DP(\alpha_h, B_h)$ following Eq.* (1)*; Assign one edge to each label y_i that links to h_{z_i}, where $z_i \sim \boldsymbol{\pi}_h$ following Eq.* (2)*.*
2) *Generate infinite number of potential functions ('types' of edges) $\Psi = \{\psi_1, \ldots, \psi_\infty\}$ and a distribution $\boldsymbol{\pi}_\psi$ from a DP process $DP(\alpha_\psi, B_\psi)$ following Eq.* (1)*; Assign*

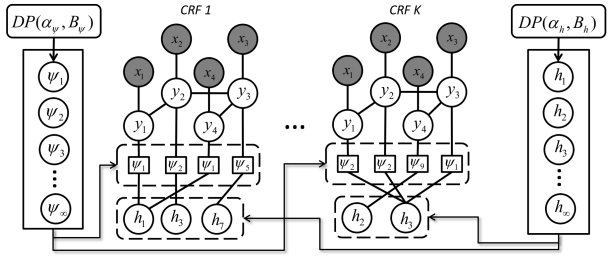


Fig. 4: Graphical representations of our infinite latent CRF (ILCRF).

one potential function ψ_{ω_i} to each edge (y_i, h_{z_i}), where $\omega_i \sim \boldsymbol{\pi}_\psi$ following Eq. (2). ■

We will illustrate the process using Figure 4. Consider first sampled CRF ('CRF-1' in the figure) with four visible nodes y_i ($i = 1 \dots 4$). In the first step, y_1 is connected to h_1, y_2 to h_3, y_3 to h_7 and y_4 to h_1 again. This is because z_i's ($i = 1 \dots 4$) are sampled as $(1, 3, 7, 1)$ from $DP(\alpha_h, B_h)$. Since only h_1, h_3 and h_7 are active, we draw their values from $DP(\alpha_h, B_h)$. Thus, we get a CRF with three latent nodes $\{h_1, h_3, h_7\}$. In the second step, the potential function of edge (y_1, h_1) is assigned to ψ_1, (y_2, h_3) to ψ_2, (y_3, h_7) to ψ_5 and (y_4, h_1) to ψ_1. This is because ω_i's are sampled as $(1, 2, 5, 1)$ from $DP(\alpha_\psi, B_\psi)$. Since, only (ψ_1, ψ_2, ψ_5) are active, we have three edge types in this CRF. We draw their parameters from $DP(\alpha_\psi, B_\psi)$. Repeating this procedure may generate different latent CRFs such as *'CRF-K'* which has two different latent nodes and three different edge types. In the end, their mixture forms the ILCRF.

The structure of labels (edges between y_i's) is defined by E_Y and is shared across all the sampled CRFs, but the process above generates different structures between the latent variables and nodes $\mathcal{Y}$. We use G_ℓ to donate the graph structure of ℓ^{th} sampled CRF. The overall likelihood of a ILCRF is given by,

$$\begin{aligned} P(\mathcal{Y}|\mathcal{X}) &= \int\int \sum_{G_\ell} P(\mathcal{Y}, G_\ell, \mathcal{H}, \Psi|\mathcal{X})\, d\mathcal{H} d\Psi \qquad (3) \\ &= \int\int \underbrace{P(\mathcal{H}|\alpha_h, B_h)}_{\text{DP prior for } \mathcal{H}} \underbrace{P(\Psi|\alpha_\psi, B_\psi)}_{\text{DP prior for } \Psi} \times \\ &\quad \Big(\sum_{G_\ell} \underbrace{P(G_\ell)}_{\substack{\text{prob. of the} \\ \text{CRF's structure}}} \underbrace{P(\mathcal{Y}|\mathcal{X}, G_\ell, \mathcal{H}, \Psi)}_{\substack{\text{conditional prob.} \\ \text{of the CRF}}} \Big) d\mathcal{H} d\Psi, \end{aligned}$$

where,

$$P(G_\ell) = \left(\prod_{i=1}^{n} \pi_h^{(z_i)} \right) \left(\prod_{i=1}^{n} \pi_\psi^{(\omega_i)} \right),$$

$$P(\mathcal{Y}|\mathcal{X}, G_\ell, \mathcal{H}, \Psi) \propto \left(\prod_{i=1}^{n} \psi_{\omega_i}(y_i, h_{z_i}) \right) \left(\prod_{(y_i, y_j) \in E} \psi(y_i, y_j) \right)$$

Exact computation on this likelihood is prohibitive in practice. We therefore present learning and inference methods based on Gibbs sampling in the following.

C. Gibbs Sampling for Learning and Inference

Gibbs sampling states that, if we sample latent CRFs, including the edge/structure G, the value of latent nodes $\mathcal{H}$ and the edge types Ψ, from their posterior distributions, then the samples approach the joint distribution $P(\mathcal{Y}, G_\ell, \mathcal{H}, \Psi|\mathcal{X})$. And this can be further used to estimate $P(\mathcal{Y}|\mathcal{X})$ in (3) and to infer the most likely values of $\mathcal{Y}$.

We present the posterior distributions below, modified from the Chinese restaurant process [23, 34] for classic DP mixture models.

- Sample the graph structure, i.e., one edge for each y_i to one latent node:[1]

$$z_i = z \propto \begin{cases} \frac{n^h_{-i,z}}{n+m-1+\alpha_h} \psi_{\omega_i}(y_i, h_z) & n^h_{-i,z} \geq 0, \\ \frac{\alpha_h/m}{n+m-1+\alpha_h} \psi_{\omega_i}(y_i, h_z) & \text{otherwise} \end{cases} \qquad (4)$$

- Sample values for each latent node in the graph:

$$h_k = h \propto B_h(h) \times \prod_{i: z_i = k} \psi_{\omega_i}(y_i, h) \qquad (5)$$

- Assign the type of potential functions to each edge:[2]

$$\omega_i = \omega \propto \begin{cases} \frac{n^\psi_{-i,\omega}}{n+m-1+\alpha_\psi} \psi_\omega(y_i, h_{z_i}) & n^\psi_{-i,\omega} \geq 0, \\ \frac{\alpha_\psi/m}{n+m-1+\alpha_\psi} \psi_{\omega_i}(y_i, h_{z_i}) & \text{otherwise} \end{cases} \qquad (6)$$

- Sample the parameters of each selected potential function:

$$\psi_k = \psi \propto B_\psi(\psi) \times \prod_{i: \omega_i = k} \psi_\omega(y_i, h_{z_i}) \qquad (7)$$

- Sample labels:

$$y_i = y \propto \psi_{\omega_i}(y, h_{z_i}) \times \prod_{(y_i, y_j) \in E} \psi(y_i, y_j) \qquad (8)$$

As for learning the E_Y, when labels are given in the training data, E_Y is independent with latent variables $\mathcal{H}$ (if the partition function is ignored), and therefore can be learned separately. In the next section, we will first describe how we model the humans as latent variables, and then specify the details of the different terms in ILCRF for our two applications.

IV. ILCRF for Scene Arrangement

In this section, we apply ILCRFs to the application of arranging objects in 3D scenes. The goal is to find the appropriate locations and orientations for placing given objects in a 3D scene (see Figure 2). In the following, we describe how to model hidden humans in a scene using ILCRF as well as learning and inference for this particular application, followed by related work in scene modeling.

[1] The posterior distribution of an variable is proportional to its prior and to its likelihood. In the case of z_i, it means that the probability of linking an edge from y_i to h_z is determined by: 1) the likelihood of this edge, given by $\psi_{\omega_i}(y_i, h_z)$; 2) the number of other subjects choosing the same latent node, i.e., $n^h_{-i,z}$ where $n^h_{-i,z} = I\{z_j = z, j \neq i\}$. In addition, the chance of selecting a new latent node is given by α_h/m out of m latent nodes sampled from B_h. (See [23] for more details).

[2] Similar to (4), the probability of choosing ψ_ω is proportional to the number of other edges choosing the same function ($n^\psi_{-i,\omega}$) and the likelihood of this edge using this function.

A. Modeling Humans in a Scene

We model possible human poses as latent nodes $\mathcal{H}$. A human pose is specified by its pose, location and orientation. Following [14], we use six types covering different sitting and standing poses.

We model object affordances as the potential functions Ψ. We use the spatial relationship between a human pose and the object to represents its affordance. It is defined as a product of several terms, each of which captures one type of spatial relation: Euclidean distance, relative angle, orientation difference,[3] and height (vertical) distance. (See [14] for details.) Thus, an affordance is defined by the parameters used in these terms. Sampling an affordance, such as ψ_k in (7), is actually sampling the parameters. See Figure 2 for the top-view of the affordances of some objects.

B. Learning and Inference

Following Defn. 3, we define $y_i \in \mathcal{Y}$ as the placement (location and orientation) of an object and $x_i \in \mathcal{X}$ as its given object class. The edges between the visible nodes $\mathcal{Y}$ model the object-object spatial relationships.[4]

During training, our goal is to learn the object affordances (i.e., a set of potential functions ψ in the ILCRF). We perform sampling on the human-object edges, human poses and object affordances, given placements $\mathcal{Y}$ and edge types ω_i, according to Eq. (4), (5) and (7). (Here, since x_i as the object class label is given, we set $\omega_i = x_i$ in this application.) The object-object structure, E_Y is learned based on object co-occurence, as computed from the training data.

During testing on a new scene, our goal is to predict placements y_i, given objects $\mathcal{X}$. We perform inference by sampling the human-object edges, human poses and placements, using the learned object affordances (see Eq. (4), (5) and (8)). In order to predict the most likely placement for object i, we choose the placement area sampled most because that represents the highest probability.

C. Related Work: Scene Modeling

To our best knowledge, there is little work about arranging/placing objects in robotics (e.g., [7, 30, 11, 16, 15]), and none of these works consider reasonable arrangements for *human usage*. In recent work, Jiang et al. [14], Jiang and Saxena [13] considered hallucinating humans for object placements and later applied similar idea to the task of scene labeling [17]. However, their method did not model human-object and object-object relationships in a joint model. They first computed a distribution of arrangements using human context only, and another using object context only. Then they linearly combined the two as the final distribution. Unlike this heuristic approach, we propose a unified model that infer the arrangement based on joint distribution of the two (see (8)). We compare our ILCRF algorithm to their approach in our experiments.

There are other recent works applying object affordances in tasks of predicting human workspaces [9]. When humans are observed, affordances can be used to predict 3D geometry [6], improve human robot interactions [24], detect and anticipate human activity [21, 19, 20]. While these works focus on different problems and require the presence of humans, they all demonstrate the advantages of considering object affordances.

[3]Only when the orientation of the object is defined and accessible.

[4]It is defined as a multi-variate Gaussian distribution of the location and orientation difference between the two objects.

TABLE I: Results of arranging partially-filled scenes and arranging empty scenes in synthetic dataset, evaluated by the location and height difference to the labeled arrangements.

Algorithms	partially-filled scenes		empty scenes	
	location (m)	height (m)	location (m)	height (m)
Chance	2.35±0.23	0.41±0.04	2.31±0.23	0.42±0.05
Obj. [14]	1.71±0.23	0.13±0.02	2.33±0.17	0.44±0.04
CRF	1.69±0.05	0.12±0.01	2.17±0.07	0.39±0.01
ILCRF-H [14]	1.48±0.18	0.11±0.01	1.65±0.20	0.12±0.01
Human+obj [14]	1.44±0.18	**0.09** ±0.01	1.63±0.19	0.11±0.01
ILCRF-Aff.	1.59±0.06	0.14±0.01	1.60±0.06	0.15±0.01
ILCRF-NSH	1.64±0.05	0.15±0.01	1.77±0.06	0.16±0.01
FLCRF	1.55±0.06	0.12±0.01	1.63±0.06	0.14±0.01
ILCRF	**1.33±0.19**	**0.09±0.01**	**1.52±0.06**	**0.10±0.01**
Obj. (+furniture)	1.63±0.05	0.15±0.01	1.80±0.05	0.20±0.01
CRF (+furniture)	1.62±0.05	0.15±0.01	1.78±0.05	0.16±0.01
Human+obj (+furniture)	1.46±0.06	0.11±0.01	1.57±0.06	0.15±0.01
ILCRF (+furniture)	**1.28±0.06**	**0.10±0.01**	**1.43±0.06**	**0.10±0.01**

TABLE II: Results on arranging five real point-cloud scenes (3 offices & 2 apartments). **Co**: % of semantically correct placements, **Sc**: average score (0-5).

	office1		office2		office3		apt1		apt2		AVG	
	Co	Sc	Co	Sc	Co	Sc	Co	Sc	Co	Sc	Co	Sc
Obj.	100	4.5	100	3.0	45.0	1.0	20.0	1.8	75.0	3.3	68.0	2.7
ILCRF-H	100	**5.0**	100	4.3	91.0	4.0	74.0	3.5	88.0	4.3	90.0	4.2
Human+obj	100	4.8	100	4.5	92.0	4.5	89.0	4.1	81.0	3.5	92.0	4.3
ILCRF	**100**	**5.0**	**100**	**4.6**	**94.0**	**4.6**	**90.0**	**4.1**	**90.0**	**4.4**	**94.8**	**4.5**

V. Experiments

In our application, the scenes (including objects/furnitures) are perceived as point-clouds (Fig. 11), either generated from 3D models in synthetic datasets or obtained using Microsoft Kinect camera in real datasets.

Dataset. We use the same two datasets as in [14, 15]: 1) a synthetic dataset consisting of 20 rooms (living rooms, kitchens and offices) and 47 objects from 19 categories (book, clean tool, laptop, monitor, keyboard, mouse, pen, decoration, dishware, pan, cushion, TV, desk light, floor light, utensil, food, shoe, remote control, and phone); 2) five real offices/apartments each of which is asked to arrange 4, 18, 18, 21 and 18 number of objects.

Experimental setup. For the synthetic dataset, we conduct 5-fold cross validation on 20 rooms so that the test rooms are new to the algorithms. We also consider two different testing scenarios, same as in [14]: 1) arranging *partially-filled* rooms by placing only one type of objects; 2) arranging *empty* rooms (with only furnitures) by placing multiple types of objects. For the real dataset, we test on the five empty rooms using learned object affordances from the synthetic dataset.

Algorithms. We compare all the following methods:

1) Chance. Objects are placed randomly in the room.

2) Obj. It uses heuristic object-object spatial relations to infer placements in sequence (not jointly), same as [14].

3) CRF, a ILCRF with only object-object edges, without latent human nodes.

4) ILCRF-H, a ILCRF with only human-object edges, without considering object relations (referred as 'DP' in [14]).

5) Human+obj. After getting the inferred distributions of arrangements $\mathcal{Y}$ from Obj. and ILCRF-H separately, it linearly

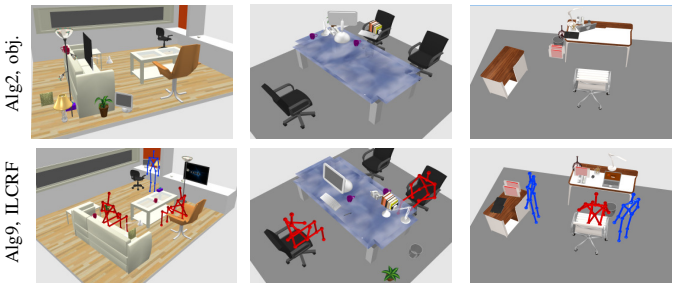


Fig. 5: Results of arranging empty rooms by using object-object relationships only (Alg2, top row) and by ILCRFs (bottom row).

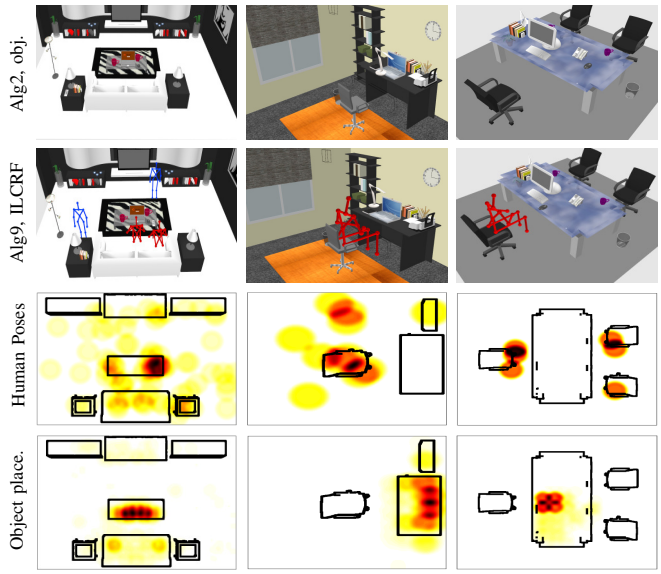


Fig. 6: Results of arranging a remote control (left), a desklight (middle) and a mouse (right), in partially-filled rooms by Alg2 (top row) and by ILCRFs (second row), We also show the top view of distribution of sampled human poses (third row) and object locations (last row) in heatmaps.

combines the two and find the maximum [14]. Our ILCRF, on the other hand, incorporate the two relationships during the inference, not after.

6) ILCRF-Aff, a ILCRF with only one type of edge, i.e., one shared affordance across all object classes. Having a universal potential function is often assumed in many CRF applications. However, it may not be appropriate for modeling object affordances.

7) ILCRF-NSH, a ILCRF with with non-sharing latent human node. It assigns one human node to each object. This model resembles *hidden CRFs* (Fig. 3). While the arrangement of an object can still be affected by its relation to possible human poses, it cannot capture phenomena of objects sharing the same human pose, such as a monitor and a keyboard. Sharing latent nodes is achieved in ILCRF by the clustering effect inherited from DPs.

8) FLCRF, a ILCRF with fixed number of latent nodes, shared across all scenes. It requires a good estimate on the number of human poses, and the optimal number may vary for rooms of different types or sizes.

9) ILCRF, our full ILCRF model.

Results. Table I presents the results on the synthetic dataset, where the predicted arrangements are evaluated by two metrics, same as in [14]: location difference and height difference (in meters) to the labeled arrangements. We also experiment using furniture information to improve arrangements. More details, including results for each object category can be found in the supplementary material. Results on the real dataset are presented in Table II, evaluated by the percentage of predicted locations are semantically correct and a score of the overall arrangement between 0 and 5, labeled by two human subjects that are not associated with this project.

In the following, we analyze the results in order to study our approach conceptually and algorithmically in that whether: 1) human-object and object-object relations are beneficial, and 2) being able to handle changeable numbers of human poses/object affordances and having a mixture of CRFs is necessary for capturing those relations.

Q1: How beneficial are the latent human poses and object affordances? From Table I we can see the performance gain quantitatively in the comparison of our full ILCRF model against ILCRF with only object edges (CRF) (and also against Obj. using the heuristic object context reported in [14]). On average, the location and height difference are reduced from 1.69m (2.17m) and .12m (.39m) to 1.33m (1.52m) and .09m (.10m), in arranging partially filled (empty) scenes. Even methods that use non-sharing skeletons (ILCRF-NSH) and finite skeletons (FLCRF) achieve better results than CRF.

We now compare some predicted arrangements visually. When arranging empty rooms (Fig. 5), only using object relationships (CRF) performs extremely poor in terms of objects crowded together and being located randomly in a scenes. This is because reasonable arrangements cannot be thoroughly explained by the object-object spatial relationships alone. For example, in the first scene, the floorlight often appears near the TV does not mean a TV should be next to it when the light is in the back of a room. In the second scene, although CRF captures that office supplies are always close to each other, it cannot capture that they should be on the table facing chairs. On the other hand, our ILCRF is able to successfully identify reasonable human workspace and use it to arrangement the objects accordingly.

When arranging one object in a scene with other objects (Fig. 6), the overcrowding of objects in CRF is alleviated, but some placements are inconvenient for human to access, such as the desk light is too close to the chair (second scene) and the mouse is far away from the chair (third scene). ILCRFs address this issue by using proper object affordance with respect to imaginary human poses. For example, most sampled human poses are on the couch/chair, and therefore most samples of the desklight are on the other end of the table and most samples of the mouse are close to the chair.

Q2: Why do we need handle unknown number of human poses? The advantage of using DP mixture models in ILCRF is being able to determine the number of human poses from the data instead of guessing manually. We investigate this in in Fig. 7. We compare ILCRF with the FLCRF where the

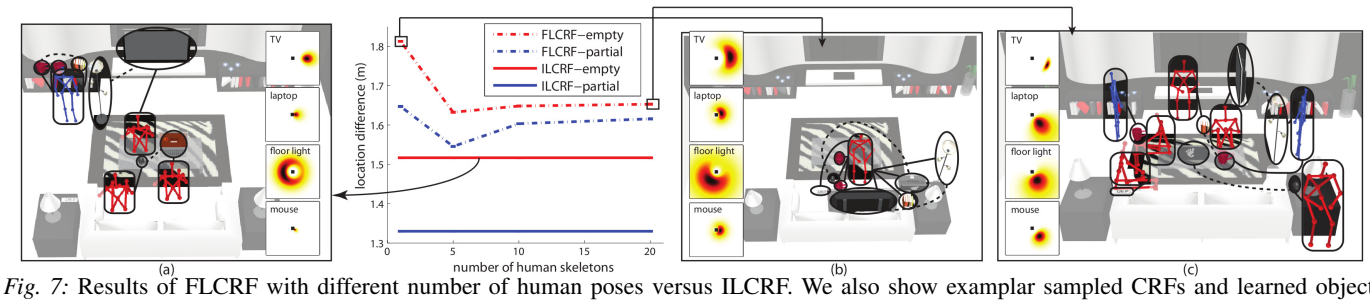


Fig. 7: Results of FLCRF with different number of human poses versus ILCRF. We also show examplar sampled CRFs and learned object affordances (in top-view heatmaps) by different methods.

number of human poses varyies from 1 to 20.

While having five poses in FLCRF gives the best result, it is still outperformed by ILCRF. This is because scenes of different sizes and functions prefer different number of skeletons. If we force all scenes using only one human pose, the learned object affordances will have large variances because all objects in the scene attempting to relate to one human, e.g., in Fig. 7-(b). If we force all scenes using a large number of human poses, say 20 per scene, the model will overfit in each scene and leading to meaningless affordances, e.g., Fig. 7-(c). Therefore, having the correct number of latent human nodes in CRFs is crucial for learning good object affordances as well as for inferring reasonable arrangements across diverse scenes (Fig. 7-a).

Q3: How sensitive is ILCRF to the number of human poses? The parameter α_h in ILCRF controls the probability of selecting a new human pose and thus can be viewed as a counterpart of K (the fixed number of human poses) in FLCRF. However, unlike FLCRF, ILCRF is much less sensitive to this parameter, as shown in Fig. 8 where its performance does not vary much for α_h from 0.1 to 10^4. Therefore, ILCRF does not rely on either informative prior knowledge or a careful hand-picked value of α_h to achieve high performance.

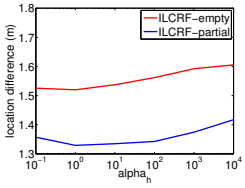


Fig. 8: The average performance of ILCRF with different hyper-parameter α_h.

Q4: Why do we need model different object affordances? Allowing diversity in object affordances is as important as that in human poses. We verify it by forcing all objects sharing the same affordance (ILCRF-Aff), shown in Fig. 9. In comparison to ILCRF, the performance drops more for partially-filled scenes than for empty scenes. This is because, in partially-filled scenes, sampled human poses are more accurate so that the performance largely depends on the correct human-object relationships. However, ILCRF-Aff still performs better than not using affordance at all, since the learned affordance still fits some objects.

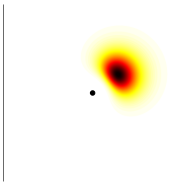


Fig. 9: Single affordance learned by ILCRF-Aff.

Q5: Why do we need a mixture of CRFs? ILCRFs is a mixture of unknown number of latent CRFs. However, we can control the number in the sampling to approximately investigate the effect of having multiple CRFs for modeling a scene.[5] Results are shown in Fig. 10. We can see that using multiple CRFs perform much better than a single CRF. However, they are all beat by ILCRFs where the number of CRFs is determined from data.

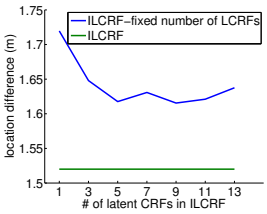


Fig. 10: Results of ILCRFs with different fixed number of latent CRFs, for empty scenes.

Q6: Why do we need to model object- and human-context jointly? Table I shows that combining the two types of context (Human+obj. and ILCRF) performs significantly better than individuals (CRF and ILCRF-H). While in Human+obj., the arrangements are inferred by the two context separately and then combined, ILCRFs combine the two during sampling (see (8)). This can make objects that have strong correlations more likely to be assigned to same human poses and thus are more likely arranged together. For example, in 77.4% of samples by ILCRF assign the keyboard and monitor to the same human pose, while only 65.8% of samples do so in Human+obj.

Q7: Can we exploit more environment knowledge? In practice, furniture in a room is often a strong cue for locating objects. To utilize the information of furniture, we model each piece of furniture as an existing object node so that when learning the object-object structure, the furniture-object relationships are also learned. We compare algorithms that use object context: Obj., CRF, Human+obj. and ILCRF (Table I).

We found that the performance gain of using furniture on empty scenes are more significant than that on partially-filled scenes. This is because in partially-filled scenes, existing objects may already provide enough object-context. In empty scenes, using furniture is especially helpful for objects such as TV, cushion and remote controls (see the category-wise results in the supplementary material). However, adding furniture sometimes even hurts the result, such as for food and phone. We conjecture this may due to some imperfectly learned object-object relationships, and this is an area of future work.

Robotic Experiment. We apply the ILCRF to our Kodiak PR2 robot to perform the scene arrangement in practice. As an example, Fig. 1 shows: Given a room (perceived in point-clouds), our robot first hallucinates human poses and labels objects in the scene [17]. After it detects the monitor and the

[5]We do so by updating the CRF structure in (4) for a limited number of times while sampling other variables regularly. The samples from the last L updates, estimate the result of having a mixture of L CRFs.

Fig. 11: From the point-clouds of given scenes (top), our robot uses ILCRF to infer possible human poses (bottom, shown in red heatmaps) and possible placements for a cushion, mouse and mug (bottom, from left, shown in blue heatmaps).

ground (Fig. 1-a), it uses our scene arrangement algorithm to infer possible human poses and possible locations for the mouse (Fig. 1-b,c). It finally places the mouse at the most likely location (Fig. 1-d).

We test our system on a small set of objects (a cushion, mouse and mug) in a given scene (Fig. 11). We visualize the sampled human poses and object locations in red and blue heatmaps. We can see that most sampled humans are sitting on the couch, bean bag or chair, and the most likely location for the cushion is on the couch and the desk. To see PR2 arranging the scene in action (along with code and data), please visit: `http://pr.cs.cornell.edu/hallucinatinghumans`

VI. Conclusion

In this paper, we considered a challenging problem of robotic scene arrangement, which requires an algorithm that can handle: 1) unknown number of latent nodes (for potential human poses), 2) unknown number of edge types (for human-object interactions), and 3) a mixture of different CRFs (for the whole scene). We therefore presented a new algorithm, called Infinite Latent Conditional Random Fields (ILCRFs), together with learning and inference algorithms. Through extensive experiments and thorough analyses, we not only showed that our ILCRF algorithm outperforms the state-of-the-art results, but we also verified that modeling latent human poses and their relationships to objects are crucial to reason our environment. Finally, we also implemented our algorithm on a robot. It correctly inferred potential human poses and object arrangements in real scenes.

Acknowledgments

This research was funded by Microsoft Faculty Fellowship and NSF Career Award to Saxena.

References

[1] A. Anand, H. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3d point clouds. *IJRR*, 32(1):19–34, 2012.

[2] A. Anandkumar, D. Hsu, F. Huang, and S. Kakade. Learning mixtures of tree graphical models. In *NIPS*, 2012.

[3] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen. the infinite hidden markov model. *NIPS*, 2002.

[4] K. Bousmalis, L.P. Morency, S. Zafeiriou, and M. Pantic. A discriminative nonparametric bayesian model: Infinite hidden conditional random fields. In *NIPS Workshop on Bayesian Nonparametrics*, 2011.

[5] J.K. Bradley and C. Guestrin. Learning tree conditional random fields. In *ICML*, 2010.

[6] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012.

[7] A. Edsinger and C.C. Kemp. Manipulation in human environments. In *Humanoid Robots*, 2006.

[8] T.L. Griffiths and Z. Ghahramani. The indian buffet process: An introduction and review. *JMLR*, 12:1185–1224, 2011.

[9] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[10] K. Ickstadt, B. Bornkamp, M. Grzegorczyk, J. Wieczorek, M. Sheriff, H. Grecco, and E. Zamir. Nonparametric bayesian networks. *Bayesian Statistics*, 9:283–316, 2010.

[11] D. Jain, L. Mosenlechner, and M. Beetz. Equipping robot control programs with first-order probabilistic reasoning capabilities. In *ICRA*, 2009.

[12] J. Jancsary, S. Nowozin, and C. Rother. Non-parametric crfs for image labeling. In *NIPS Workshop Modern Nonparametric Methods Mach. Learn.*, 2012.

[13] Y. Jiang and A. Saxena. Hallucinating humans for learning robotic placement of objects. In *ISER*, 2012.

[14] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. In *ICML*, 2012.

[15] Y. Jiang, M. Lim, C. Zheng, and A. Saxena. Learning to place new objects in a scene. *IJRR*, 2012.

[16] Y. Jiang, C. Zheng, M. Lim, and A. Saxena. Learning to place new objects. In *ICRA*, 2012.

[17] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, 2013.

[18] H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011.

[19] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. In *RSS*, 2013.

[20] H. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML*, 2013.

[21] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.

[22] J. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[23] R.M. Neal. Markov chain sampling methods for dirichlet process mixture models. *J comp graph statistics*, 9(2):249–265, 2000.

[24] A.K. Pandey and R. Alami. Taskability graph: Towards analyzing effort based agent-agent affordances. In *RO-MAN, IEEE*, 2012.

[25] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *NIPS*. Citeseer, 2004.

[26] A. Quattoni, S. Wang, L.P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *PAMI*, 29(10):1848–1852, 2007.

[27] A. Rodriguez, A. Lenkoski, and A. Dobra. Sparse covariance estimation in heterogeneous samples. *Elec. Journal of Stat.*, 5:981–1014, 2011.

[28] A. Saxena, S.H. Chung, and A. Ng. Learning depth from single monocular images. In *NIPS 18*, 2005.

[29] P. Schnitzspan, S. Roth, and B. Schiele. Automatic discovery of meaningful object parts with latent crfs. In *CVPR*, 2010.

[30] M. Schuster, J. Okerman, H. Nguyen, J. Rehg, and C. Kemp. Perceiving clutter and surfaces for object placement in indoor environments. In *Humanoid Robots*, 2010.

[31] D. Shahaf, A. Chechetka, and C. Guestrin. Learning thin junction trees via graph cuts. *AISTATS*, 2009.

[32] N. Srebro. Maximum likelihood bounded tree-width markov networks. In *UAI*, 2001.

[33] C. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *ICML*, 2004.

[34] Y. W. Teh. Dirichlet process. *Encyc. of Mach. Learn.*, 2010.

[35] J. Van Gael, Y. W. Teh, and Z. Ghahramani. The infinite factorial hidden markov model. In *NIPS*, 2008.

[36] S.B. Wang, A. Quattoni, L.P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, 2006.