Probabilistic Pick and Place Planning instead of Pick then Place Planning

Mohanraj Devendran Shanthi[1] and Tucker Hermans[1,2]

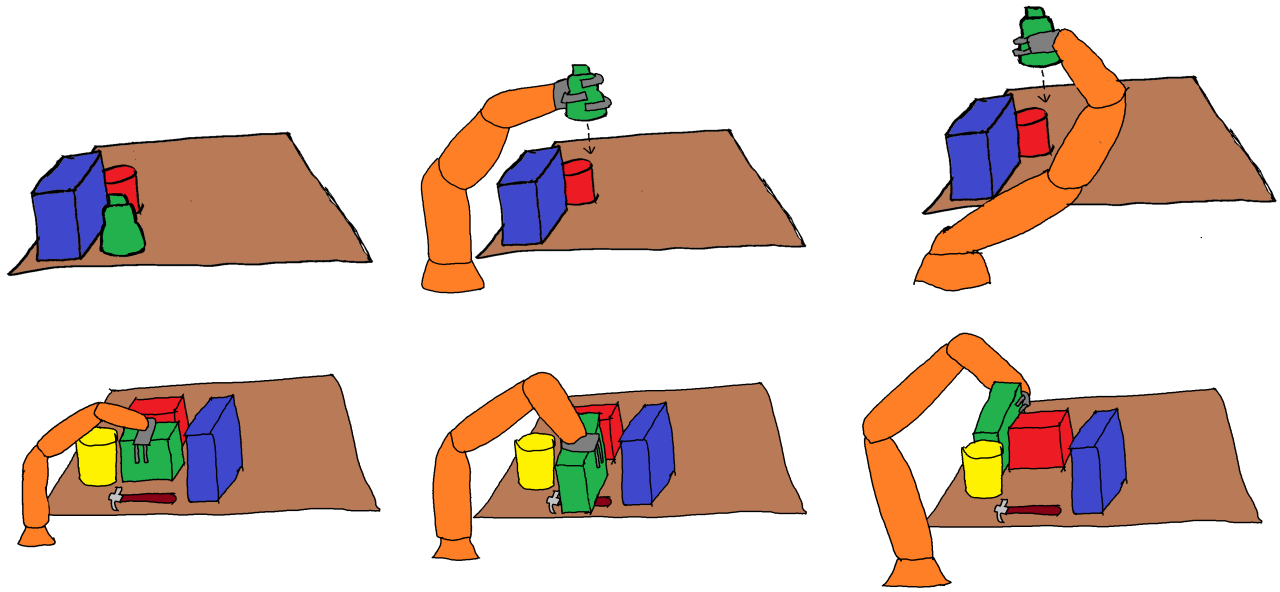


Fig. 1: Importance of grasp configuration and placement configuration for the pick and place task success. **Top row:** Effect of grasp on pick and place success. **Top-left:** Target object configuration. **Top-center:** Good grasp incompatible with the placement. **Top-right:** Good grasp compatible with placement. **Bottom row:** Effect of placement configuration on pick and place success. **Bottom-left:** Good placement configuration for grasp A. **Bottom-center** Incompatible placement configuration. **Bottom-right** Good placement configuration for grasp B.

I. Introduction

Pick and place constitutes a fundamental building block of robotic automation. With a successful place action requiring a successful pick action, we could think of the success of a pick and place task to be conditioned on the success of the pick task in a probabilistic sense. Hence, there is much merit in considering the pick and place problem as maximization of the pick and place probabilities jointly. Considerable work examines grasping objects [1,2], with contemporary methods capable of grasping novel objects with high success [3–5]. These recent successes tend to model the problem as probabilistic inference using some learned neural network model. This line of thinking enables us to extend these methods to a joint pick and place pipeline.

Research focused on object placement, though not as extensive as grasping, investigates various aspects including stability of placements [6,7], semantic placement [8,9], and multi-object rearrangement [10,11]. Though pick and place naturally go hand-in-hand, most research investigates the two highly related sub-tasks individually.

Treating the problems independently ignores a number of important issues. In particular, while grasp success is necessary for successful placement, it is not sufficient to guarantee it. A grasp configuration might succeed in lifting an object, but could end up contributing to placement failure if the robot collides with other objects in the scene during placement as shown in Fig. 1. Likewise, if one ensures that the previously planned grasp does not collide with objects during placement, it might do so at the expense of object instability at placement or reachability by the robot arm. Thus causing either placement planning or execution failure.

Those works that have examined tackling pick and place jointly, restrict themselves in someway, making simplifying assumptions not needed in modern grasp planners. These simplifications include requiring full geometry of the object and environment in the form of meshes [12], a restricted class of known object categories [13], or restricting the planner to use a fixed subset (e.g. overhead) of grasps [14] or simplified grippers [15]. In contrast, we examine the problem of joint pick and place planning given only partial view point clouds of the object and environment. This includes the case of grasping and placing previously unseen objects. Further, we plan over arbitrary grasps from the full continuous space of feasible robot configurations, as done in recent grasping work [3].

In this paper we formalize the joint pick and place task as a probabilistic inference problem (Sec. II). Our framework enables us to jointly solve for both the optimal placement location of the given object in clutter and a corresponding

[1]Robotics Center, University of Utah, Salt Lake City, UT, USA; [2]NVIDIA, Seattle, WA, USA. Email: mohanraj.devendranshanthi, tucker.hermans@utah.edu This work is supported by DARPA under grant N66001-19-2-4035 and by NSF Award #1846341.

grasp configuration suitable with the placement. We do so using only sensor information of the scene enabling our approach to work with novel objects. This joint solving approach makes the grasp and placement configurations to be flexible to each other's effects by means of propagating gradients, this ensures the planned grasp and placement configurations are compatible with each other. We use a state of the art grasp learning approach to encode the grasp success likelihood [3]. Like other works using neural networks for learning [2,16–19] the ability to compute gradients through the model allows us to use it within an efficient gradient-based solver for inference. We detail our proposed solution in Sec. III.

We validate our approach on a physical robot with a multi-fingered hand by comparing against baselines that treat the individual pick and place planning as sequential, non-interacting problems. Our results in Sec. IV show our approach outperforms these baseline with improved success rates and optimality in terms of final placement probabilities.

We make the following contributions.

- Present a framework for reasoning about pick and place planning jointly
- Provide a concrete implementation of this using a learned grasp classifier to encode grasp success in the objective
- We validate the previously proposed grasp learning method of [3] on a different gripper, further validating the approach.

II. Pick and Place as Joint Inference

Let O be an object to be placed in a cluttered environment E, with partial-view depth images Z_O and Z_E respectively. The grasp configuration $\boldsymbol{\theta}_g = [\boldsymbol{x}_g, \boldsymbol{q}_g^h]$ is a vector including the robot palm pose $\boldsymbol{x}_g \in SE(3)$ and preshape joint angles of the gripper's fingers $\boldsymbol{q}_g^h \in \mathcal{Q}_h$. The placement configuration $\boldsymbol{x}_p \in SE(3)$ defines the 6-DOF pose of where the centroid of object point cloud Z_O should be once placed. We can then define the probability of successfully grasping the object as:

$$F\left(\boldsymbol{\theta}_g; Z_O, Z_E\right) = P\left(r_g{=}1 | \boldsymbol{\theta}_g, Z_O, Z_E\right) \tag{1}$$

And the probability of the place configuration x_p being successful for object O in environment E conditioned on grasp success is defined as:

$$H\left(\boldsymbol{x}_p, \boldsymbol{q}_g; Z_O, Z_E\right) = P\left(r_p{=}1 | \boldsymbol{x}_p, \boldsymbol{q}_g, Z_O, Z_E\right) \tag{2}$$

The joint probability for pick and place success is then obtained as:

$$\begin{aligned} &P\left(r_g{=}1, r_p{=}1 | \boldsymbol{\theta}_g, \boldsymbol{x}_p, Z_O, Z_E\right) \\ &\qquad = F\left(\boldsymbol{\theta}_g; Z_O\right) H\left(\boldsymbol{x}_p, \boldsymbol{q}_g; Z_O, Z_E\right) \end{aligned} \tag{3}$$

Which we visualize as a factor graph in Fig.2. We see that while the success probabilities are conditionally independent given the planning parameters, they can not be fully decoupled, requiring joint inference over pick and place parameters

We define the pick and place inference problem as finding a tuple of grasp configuration and place configurations

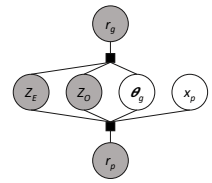


Fig. 2: Factor graph of the pick and place probability distribution. We see that while the success probability are conditionally independent given the planning parameters, they can not fully decouple, requiring joint inference over pick and place parameters.

$(\boldsymbol{\theta}_g, \boldsymbol{x}_p)$, that maximizes the joint probability defined in Eq. (3). Taking the negative log on Eq. (3), we formalize this as a constrained optimization in Eq. (4)

$$\min_{\boldsymbol{x}_p, \boldsymbol{\theta}_g, \boldsymbol{q}_g^a, \boldsymbol{q}_p^a, \tau} -\log\big(H\left(\boldsymbol{x}_p, \boldsymbol{q}, Z_O^+; Z_E\right)\big) - \log(F\left(\boldsymbol{\theta}_g; Z_O\right)) \tag{4a}$$

subject to

$$\boldsymbol{x}_p \in \mathcal{P} \tag{4b}$$

$$\boldsymbol{x}_g = \phi_h(\boldsymbol{q}_g^a); \quad \boldsymbol{x}_p = \phi_O(\boldsymbol{q}_p^a) \tag{4c}$$

$$\boldsymbol{q}_i^- \leq \boldsymbol{q}_i \leq \boldsymbol{q}_i^+ \quad \forall\, i \in \{g, p\} \tag{4d}$$

$$Z_O^+ = Z_O \cup R_G(\boldsymbol{\theta}_g) \tag{4e}$$

$$\epsilon \leq \texttt{SDF}\left(\boldsymbol{x}_p, Z_O^+\left(\boldsymbol{\theta}_g\right), Z_E\right) \tag{4f}$$

$$\tau(\boldsymbol{x}_p, \boldsymbol{x}_g) \in \Omega \tag{4g}$$

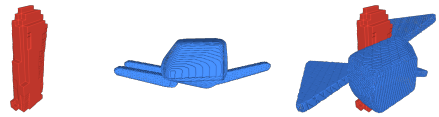
Equation (4a) defines the objective of the optimization as a log-linear combination of the placement success probability $H\left(\boldsymbol{x}_p, Z_O; Z_E\right)$ and grasp success probability $F\left(\boldsymbol{\theta}_g; Z_O\right)$. For simplicity, we restrict the search to be over 2D planar placement configurations within a predefined region, $\boldsymbol{x}_p \in \mathcal{P} \subset SE(2)$ (Eq. 4b). In doing so, we restrict the object to be placed resting on the same face as it was when picked. We could easily extend to full 3D poses given an appropriate models predicting the stability of placement orientations.

The remaining constraints ensure physical validity for successful execution, i.e., the grasp and placement must be reachable by the robot and the objects and robot should not interpenetrate. Equation (4c) encodes the arm forward kinematics for the grasp and placement, while Eq. (4d) defines the joint limits, where the superscript, i, denotes joints associated with the arm.

Equation (4e) augments the object cloud with the robot gripper geometry, $R_G(\boldsymbol{\theta}_g)$, according to the current grasp pose. We note this is a similar procedure to that in [28]. We visualize the gripper geometry augmentation in Fig. 3. Using this we define the placement collision constraint in Eq. (4f). Finally, Eq. (4g) defines that there must be a feasible, collision-free trajectory from grasp to placement.

III. Solving the Joint Pick and Place Inference

In this section we discuss our approach to instantiating and solving the problem defined by Eq. (4). We first discuss the details of the placement probability, $H(\cdot)$, which we examine in our experiments. We then briefly review the learning-based grasp method from [3] and its use as our grasp probability $F(\cdot)$. Following that we present an efficient algorithm for SDF-based collision checking built specifically for repeated placement into clutter.



(a) (b) (c)

Fig. 3: Object-grasp geometric modeling for collision checking. 3a object geometry. 3b robot geometry. 3b augmented geometry:union of object and robot geometries at grasp configuration $\boldsymbol{\theta}_g$.

A. Placement Likelihood

The placement likelihood $H(\boldsymbol{x}_p)$ defines the objective to evaluate object placement locations to achieve a specific placement task. It accepts a placement configuration in $SE(2)$ as input and outputs a scalar value $\in [0, 1]$ with values closer to 0 being less desirable for the task and values close to 1 being more desirable. While having simple analytical models for $H(\boldsymbol{x}_p)$ makes it convenient to derive gradients, we could also use differentiable, learned models.

We define a Gaussian placement likelihood that prefers placement configurations $\boldsymbol{x}_p$ close to a target pose $\boldsymbol{x}_t$ as:

$$H_{\text{target}}(\boldsymbol{x}_p; \boldsymbol{x}_t) \propto \exp\left(-\frac{1}{2}\alpha\left(\boldsymbol{x}_t - \boldsymbol{x}_p\right)^T I\left(\boldsymbol{x}_t - \boldsymbol{x}_p\right)\right) \quad (5)$$

Here, α is a hyper parameter, higher values of alpha makes the solver prefer placements closer to the target x_t at the expense of grasp success, lower values of α prefers more confident grasps. We defer more expressive placement distributions, including learned likelihoods in future work.

B. Grasp Prediction

Following the success of recent learning-based grasp planning approaches [3,29] we define our grasp cost as the probability of grasp success $F(\cdot) = P(r_g = 1|\boldsymbol{\theta}_g; Z_O)$. A neural network classifier defines the core of the the grasp prediction model $P(r_g = 1|\boldsymbol{\theta}_g; Z_O)$. This outputs a scalar value between 0 and 1 that represent the grasp success probability for the given grasp on the observed object. We learn this model as a 3D convolutional neural network classifier using the approach proposed in [3]. This takes a voxel representation of the object, converted from the point cloud, as input to and passes it through several 3D convolutional layers to predict grasp success. The only modification we make to the neural network structure is changing the grasp input model to accept the one-dimensional preshape configuration (the finger spread) instead of the higher-dimensional vector used for the dexterous hand in [3].

In addition to the classifier, the model has a mixture density network (MDN) which takes the same voxel grid as input and can generate grasp samples used for initializing the grasp planning process (Figure 5). We make use of this to generate initial samples in our planning algorithm described later in this section. For further details of the method and network architecture see [3].

C. SDF Collision Constraint Computation

To account for the collision constraints in Eq 4f, we require signed distances from the partial view points of the objects in the environment. For efficient computation we compute a discrete approximation of the SDF for each object that we can quickly update as more objects are placed into the scene. Figure 4 shows the steps in generating the SDF queries for the collision constraint.

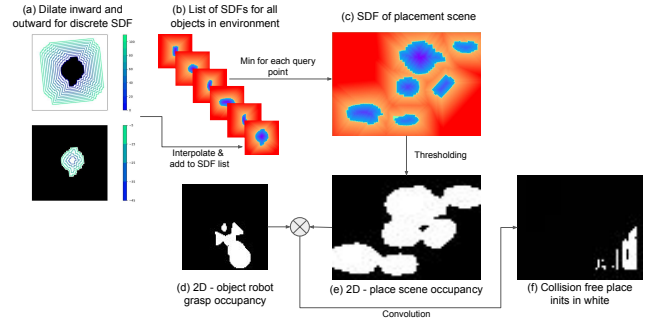


Fig. 4: Steps to generate signed-distance function for place scene and computing collision-free place initializations.

(a) (b) (c) (d)

Fig. 5: MDN prior used for initialization, showing the top and side grasp modes for the partial view point clouds of the objects a lego blocks, b cracker box, c mustard bottle and d pitcher

D. Optimization Solver

We perform MAP inference by solving the optimization problem from Eq. (4) without constraint Eq. (4g). We relax the forward kinematics Eq. (4c) and the collision SDF constraints Eq. (4f) into the objective using an Augmented Lagrangian method. We convert the constraint, $\boldsymbol{x} \in \mathcal{P}$, to bound constrain the 2D planar placement of, $\boldsymbol{x}_p$, within the table edges. We solve the resulting bound constrained problem using BFGS [30] with projections to handle the bounds on the joint angles and placement pose. We ensure Eq. (4g) when motion planning for the arm.

To handle local minima and infeasibilities, the solver is initialized with grasp configurations ${\boldsymbol{\theta}_g}^0$ sampled from the MDN prior described in III-B and shown in Figure 5. The place initialization is then obtained by convolving the 2D binary occupancy of the augmented object-robot geometry (Figure 4d) over the coarse 2D binary occupancy of the place scene (Figure 4e), which outputs collision-free place configurations (Figure 4f), these are then ranked by predicted place probabilities and filtered by kinematic feasibility, to obtain the initial place configuration ${\boldsymbol{x}_p}^0$

IV. Experiments

We validate the benefits of the joint pick and place framework by benchmarking against sequential pick and place baselines on a real KUKA LBR iiwa 14 R820 7-dof arm, with a Reflex Takktile 2 gripper, we use a Intel Realsense D455 for sensing the object to be placed.

The following baselines are considered:

1) **The pick then place approach:** Where we solve for the best grasp configuration subject to all the constraints mentioned in the approach that applies, then solving for

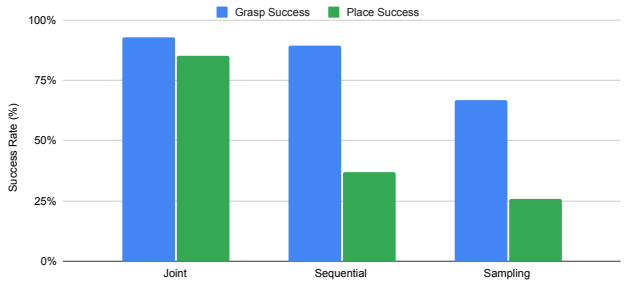


Fig. 6: Grasp and place success rates for the joint inference, sequential inference and sampling methods across 27 executions.

the placement configuration that suits the grasp. This is essentially done by solving the optimization problem described in II and III twice with the grasp and place costs individually.

2) **Sampling:** We develop a baseline that generates compatible grasp and placement configurations using Monte Carlo sampling. First we generate a set grasp configurations with high success rate from the trained grasp net, and a set of placement configurations not in collision with the environment for both object and robot using the generated sdf, then the generated grasp and placement configurations are refined locally for feasibility with other constraints.

We report the success rate, and predicted placement probabilities for each method with 27 pick and place executions with 10 different objects in different levels of clutter ranging from 4 - 7 objects in the placement scene. Figure. 6 shows grasp and place success rates, we can see that the joint method significantly out performs the baselines in terms of place success with 85% of the executions being successful while also having high grasp success comparable to the sequential inference.

Figure 7 shows the predicted placement probabilities of successful executions for each of the 27 executions for all 3 methods, the predicted probabilities of unsuccessful executions are set as 0. It can be seen from the plot that the joint inference method outperforms the baselines in terms of predicted place probabilities in most of the successful instances, with most failure instances of the joint method also failing with the baselines. We observed the leading cause of failure being the object shifting mid trajectory. Figure 8 shows example placement executions for each method considered in each place scene.

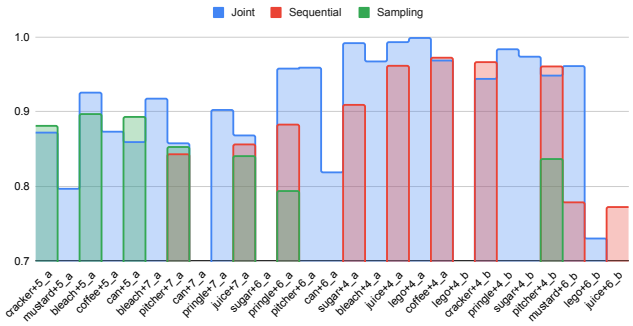
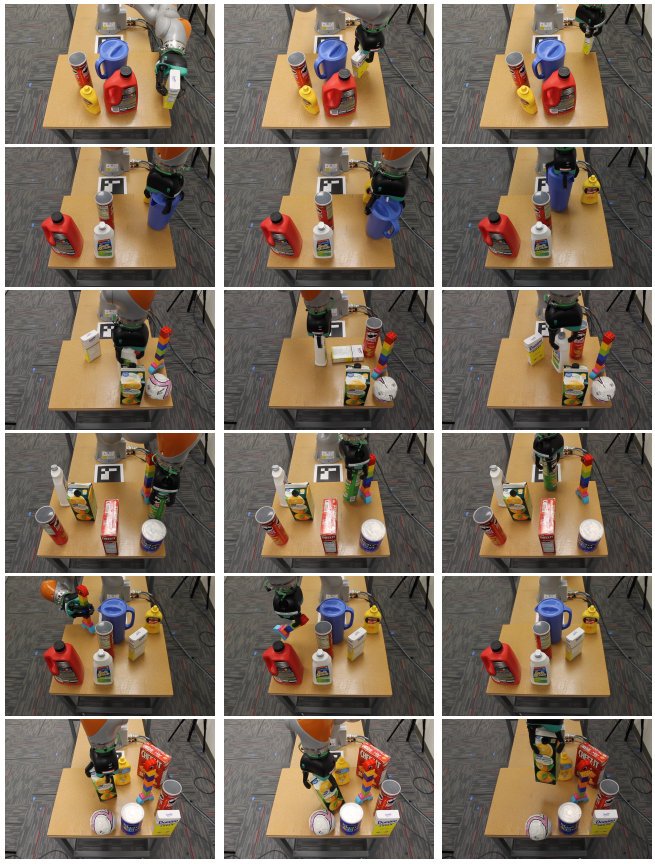


Fig. 7: Predicted placement probabilities of successful executions (placement probabilities of failed executions are set as zero)



(a) **(b)** **(c)**

Fig. 8: Example pick and place executions for each place scene. a joint inference, b sequential inference, c sampling.

V. Conclusion

We presented an approach for jointly planning a grasp for picking an unknown object jointly with a downstream placement task. By formalizing this problem as a joint inference we were able to leverage both model-based geometric and learning-based costs and constraints into a single framework. There are many opportunities for future work. Most straightforward would be learning a placement classifier, akin to the grasp classifier, in order to handle placement on non-planar surfaces. The other major place for improvement is using haptic and tactile feedback to reason about contact during placement to account for shifting in the object pose relative to the gripper during transport as well as enabling gentler and tighter placement. This could also be coupled with learning to predict how the object might move mid-trajectory after grasping. Finally, we hope to examine task-specific grasp planning for downstream use other than placement (e.g. tool use, handover).

In conclusion, our work is the first to show unified planning of a multi-fingered grasp for pick and place operations. Our results show the benefit of taking the placement location into account when planning grasps, in particular enabling higher success in placement in cluttered scenes relative to planning placements sequentially after a successful grasp.

REFERENCES

[1] R. Grupen, "Planning grasp strategies for multifingered robot hands," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 1991.

[2] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 4304–4311.

[3] Q. Lu, M. Van der Merwe, B. Sundaralingam, and T. Hermans, "Multifingered grasp planning via inference in deep neural networks: Outperforming sampling by learning differentiable models," *IEEE Robotics Automation Magazine*, vol. 27, no. 2, pp. 55–65, 2020.

[4] D. Morrison, J. Leitner, and P. Corke, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *Robotics: Science and Systems (RSS)*, 2018.

[5] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Robotics: Science and Systems (RSS)*, 2017.

[6] F. Furrer, M. Wermelinger, H. Yoshida, F. Gramazio, M. Kohler, R. Siegwart, and M. Hutter, "Autonomous robotic stone stacking with online next best object target pose planning," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2017, pp. 2350–2356.

[7] J. Baumgartl, T. Werner, P. Kaminsky, and D. Henrich, "A fast, gpu-based geometrical placement planner for unknown sensor-modelled objects and placement areas," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 1552–1559.

[8] C. Paxton, C. Xie, T. Hermans, and D. Fox, "Predicting Stable Configurations for Semantic Placement of Novel Objects," in *Conference on Robot Learning (CoRL)*, 2021.

[9] Y. Jiang, M. Lim, C. Zheng, and A. Saxena, "Learning to place new objects in a scene," *The International Journal of Robotics Research*, vol. 31, no. 9, p. 10211043, May 2012.

[10] S. Han, N. Stiffler, A. Krontiris, K. Bekris, and J. Yu, "High-quality tabletop rearrangement with overhand grasps: Hardness results and fast methods," in *Robotics: Science and Systems (RSS)*, 2017.

[11] A. Cosgun, T. Hermans, V. Emeli, and M. Stilman, "Push Planning for Object Placement on Cluttered Table Surfaces," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2011.

[12] J. A. Haustein, K. Hang, J. Stork, and D. Kragic, "Object placement planning and optimization for robot manipulators," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[13] M. Gualtieri and R. W. Platt, "Learning 6-dof grasping and pick-place using attention focus," in *Conference on Robot Learning (CoRL)*, 2018.

[14] L. Berscheid, P. Meibner, and T. Kroeger, "Self-supervised learning for precise pick-and-place without object model," *IEEE Robotics and Automation Letters*, 2020.

[15] Z. He, N. Chavan-Dafle, J. Huh, S. Song, and V. Isler, "Pick2place: Task-aware 6dof grasp estimation via object-centric perspective affordance," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 05 2023, pp. 7996–8002.

[16] B. Wu, I. Akinola, and P. K. Allen, "Pixel-attentive policy gradient for multi-fingered grasping in cluttered scenes," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[17] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Generating grasp poses for a high-dof gripper using neural networks," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[18] J. Varley, J. Weisz, J. Weiss, and P. Allen, "Generating multi-fingered robotic grasps via deep learning," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 4415–4420.

[19] M. Veres, M. Moussa, and G. W. Taylor, "Modeling grasp motor imagery through deep conditional generative models," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 757–764, 2017.

[20] J. M. Romano, K. Hsiao, G. Niemeyer, S. Chitta, and K. J. Kuchenbecker, "Human-inspired robotic grasp control with tactile sensing," *IEEE Transactions on Robotics*, vol. 27, no. 6, 2011.

[21] B. Sundaralingam, A. Lambert, A. Handa, B. Boots, T. Hermans, S. Birchfield, N. Ratliff, and D. Fox, "Robust Learning of Tactile Force Estimation through Robot Interaction," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2019.

[22] Z. Li and S. Sastry, "Task-oriented optimal grasping by multifingered robot hands," *IEEE Journal on Robotics and Automation*, vol. 4, no. 1, pp. 32–44, 1988.

[23] R. Newbury, K. He, A. Cosgun, and T. Drummond, "Learning to place objects onto flat surfaces in upright orientations," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4377–4384, 2021.

[24] A. Zeng, S. Song, K. Yu, E. Donlon, F. R. Hogan, M. Bauzá, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. J. Taylor, W. Liu, T. A. Funkhouser, and A. Rodriguez, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2018.

[25] J. Zhao, D. Troniak, and O. Kroemer, "Towards robotic assembly by predicting robust, precise and task-oriented grasps," in *Conference on Robot Learning (CoRL)*, 2020.

[26] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021.

[27] C. Mitash, R. Shome, B. Wen, A. Boularias, and K. Bekris, "Task-driven perception and manipulation for constrained placement of unknown objects," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5605–5612, 2020.

[28] F. Wang and K. Hauser, "Robot packing with known items and nondeterministic arrival order," in *Robotics: Science and Systems*, 2019.

[29] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[30] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.