Visual Imitation Learning of Task-Oriented Object Grasping and Rearrangement

Yichen Cai*, Jianfeng Gao*, Christoph Pohl, and Tamim Asfour

***Abstract*— Task-oriented object grasping and rearrangement are critical skills for robots to accomplish different real-world manipulation tasks. However, they remain challenging due to partial observations of the objects and shape variations in categorical objects. In this paper, we propose the *Multi-feature Implicit Model* (MIMO), a novel object representation that encodes *multiple spatial features* between a point and an object in an implicit neural field. Training such a model on multiple features ensures that it embeds the object shapes consistently in different aspects, thus improving its performance in object shape reconstruction from partial observation, shape similarity measure, and modeling spatial relations between objects. Based on MIMO, we propose a framework to learn task-oriented object grasping and rearrangement from single or multiple human demonstration videos. The evaluations in simulation show that our approach outperforms the state-of-the-art methods for multi- and single-view observations. Real-world experiments demonstrate the efficacy of our approach in one- and few-shot imitation learning of manipulation tasks.**

I. Introduction

Performing accurate manipulation tasks with everyday objects is an intricate problem that poses several challenges for robots. The robot must first find the optimal grasps for specific tasks and generate a suitable motion trajectory to achieve this configuration. For instance, a side grasp by the mug handle is suitable for pouring water out of a mug (see Fig. 1a), while a top grasp by the rim is more suitable when placing the mug into a container to avoid collision between the hand and the container (see Fig. 1b). Additionally, suitable pose configurations of the mug relative to the bowl and the container are needed in such an object rearrangement task.

To generate task-oriented grasps, previous works [1]–[3] have focused on training neural networks on large manually annotated datasets. Despite their performance, these approaches fail to generalize to novel objects with large shape variations. Moreover, manual annotation is costly and difficult to acquire. In contrast, visual imitation learning (VIL) approaches like [4], [5] provide efficient means to teach robots manipulation skills from human demonstrations and enable generalization to new scenarios with categorical objects. This paper focuses on the line of works that utilize neural fields, e. g., [6]–[9], which implicitly encode object

(a) Side Grasp and Pouring.

(b) Top-down Grasp and Placement.

Fig. 1: Learning task-oriented object grasping and rearrangement from human demonstration videos of manipulation tasks. We illustrate two tasks: (a) side picking a mug and pouring into a bowl; and (b) top-down picking a mug and placing it into a container. For each task, we show the RGB image, the observed point clouds (•), reconstructed object meshes (▬ ▬), extracted hand mesh (▬), grasp poses ($\mathbf{T}_g^d$), and the execution on a humanoid robot.

spatial properties. Neural fields can be trained in a self-supervised manner by exploiting an inherent bias towards object classes, thus eliminating the need for manual annotation. This bias plays an important role in establishing dense 3D correspondences across categorical objects, enabling the adaptation of object manipulation skills to previously unseen object instances. However, these approaches require multiple views of the object, which are often unavailable in real-world applications. When presented with a partial view or categorical objects with large shape variations, these approaches may yield less precise grasp or object target poses, potentially resulting in collisions or unstable placement.

To address the above-mentioned challenges, we introduce the *Multi-feature Implicit Model* (MIMO), which is designed to predict multiple spatial properties of a 3D point relative to an object. This enables our model to generate a richer descriptor space and thus more precise dense correspondences, which facilitates the accurate transfer of grasps and object target poses to new situations. MIMO can also reconstruct object shapes when only a partial observation is available, which is beneficial for coping with task constraints

*The authors contributed equally to the paper.

This work was supported by the Carl Zeiss Foundation under the project JuBot and the European Union's Horizon Europe Framework Programme under grant agreement No 101070596 (euROBIN).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. E-mails: ujfao@student.kit.edu, {jianfeng.gao, asfour}@kit.edu

defined on the hidden part of the object. Leveraging MIMO's capabilities, we propose a framework that efficiently learns and generates task-oriented grasps from single or multiple human demonstration videos. Moreover, we use an evaluation network to predict the success probability of the generated grasps and refine them if necessary.

The contributions of this paper are twofold: (1) We propose the novel *Multi-feature Implicit Model* (MIMO) that predicts multiple spatial features of a point relative to an object, which yields an informative point and pose descriptor space. It outperforms the state-of-the-art neural field methods in terms of dense correspondence, shape reconstruction, and pose transfer. The model can be trained in a self–supervised manner without relying on human annotations. (2) We integrate MIMO into visual imitation learning and propose a framework to efficiently learn, generate, and refine task-oriented grasps. We achieve one- and few-shot imitation learning and demonstrate a direct transfer of the learned manipulation tasks to categorical objects.

II. Related Work

Deep learning-based methods for grasping have made significant progress in robotics thanks to advances in implicit object representation. Explicitly modeling the relevance of manipulation skills for a given task is important for generalization to novel situations. In this regard, we focus on implicit representation through neural fields, along with recent advancements in task-relevant grasping and manipulation.

A. Neural Fields and Neural Descriptors

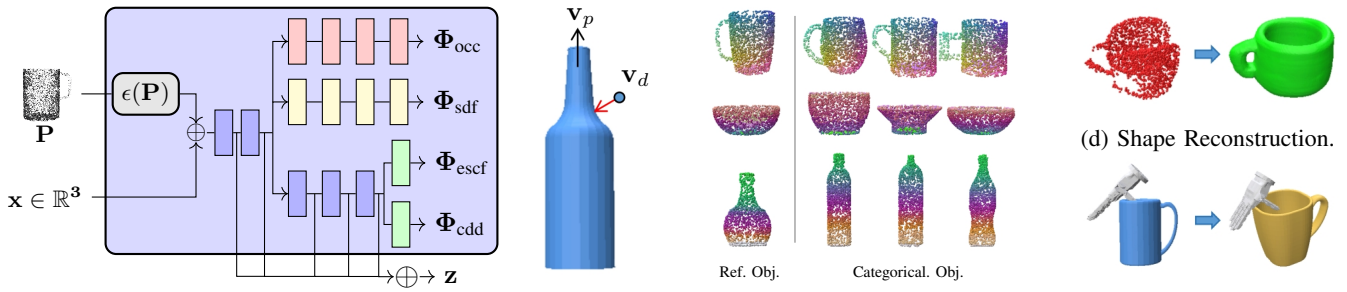
Neural-fields-based approaches involve training neural networks to learn a continuous representation by predicting the physical and spatial properties of a 3D point relative to its surroundings [6]. The learned representations, known as descriptors, are used in various tasks such as 3D reconstruction [10], [11] and manipulation [12], [13]. Leveraging dense correspondences in the descriptor space allows the transfer of manipulation skills between similar objects. Previous works [14]–[16] used Convolutional Neural Networks to obtain pixel-wise descriptors for detecting correspondences from RGB images. However, these approaches rely on visible 2D descriptors, which fail to account for task constraints on the hidden parts of the objects. To overcome this limitation, *Neural Descriptor Fields* (NDFs) [7] directly encode SE(3)-equivariant point and pose descriptors from the 3D point cloud of objects. Although a richer descriptor space was proposed in [8] by leveraging the space coverage feature (SCF) [17], it sacrificed the ability to reconstruct object shapes. Despite their performance for grasp transfer in multi-view cases with a few demonstrations (5-10), the accuracy degenerates where only a partial view or a single demonstration is available. In contrast, we train the implicit model to predict multiple spatial features of a point relative to an object, resulting in a more informative descriptor space while preserving the shape reconstruction capability. We outperform the approaches presented in [7] and [8] in tasks such as shape similarities measure and pose transfer, especially with partial view. This also leads to better performance in one-shot imitation learning of manipulation tasks.

B. Modeling Task Relevance

In the context of task-oriented grasping, it is crucial to consider the modeling of task relevance as this enables the determination of grasp poses that are most conducive to the downstream task. In previous works, semantic segmentation models have been trained to detect grasp affordance regions from either RGB images [1]–[3], [18] or 3D point clouds [19]–[22]. However, these methods often rely on large annotated datasets, necessitating time-consuming manual annotation. Furthermore, they are tailored to grasping rather than object rearrangement tasks. The former challenge is alleviated in [23], [24] via self-supervision in simulation. To address the latter, recent works focus on modeling task relevance using general 2D or 3D neural descriptors, e. g., 2D affordance regions [25], [26] and 3D affordance maps [27], [28]. These neural descriptors measure shape similarity, enabling the modeling of task relevance and facilitating the transfer of task-relevant grasps, object poses, or regions to new scenarios. However, the approaches in [25] and [26] are limited to top-down planar grasps, while multiple calibrated RGB images are required in [27] and [28] for scene reconstruction, which is time-consuming and not always feasible. In contrast, we use a novel neural pose descriptor derived from partial observations, which can be used for modeling task-oriented grasp distributions and downstream rearrangement tasks.

C. Category-Level Manipulation

Previous works, like [29]–[31], utilized semantic keypoints for transferring manipulation skills between categorical objects. However, overlooking the category-level inductive bias, these approaches necessitate extensive manual annotation for keypoint detection and careful assignment of keypoints for each task and object. To address this problem, category-level non-rigid registration [32]–[34] was proposed to reconstruct object shapes and infer object 6D poses. However, these models face difficulties in transferring to objects with large shape variations. Another line of work [7], [8] leverages category-level neural descriptors for transferring skills. However, these models assume one interacting object is known and fixed. Relational-NDF (R-NDF) [9] relaxed this limitation by manually selecting keypoints and associated local frames in task-relevant regions. However, it faces challenges in predicting precise dense correspondences with partial observation, a limitation addressed by [35] through subtasks, including pose estimation, shape reconstruction, similarity measure, and grasp transfer. Yet, each subtask demands a separate model. In contrast, we address partial observation by leveraging MIMO's capability in shape reconstruction. This enhances the precision of task relevance and knowledge transfer for object grasping and rearrangement and allows the usage of MIMO for all tasks, offering an efficient solution for manipulation tasks.



(a) Network Structure of MIMO. (b) CDD. (c) Similarity Measure. (d) Shape Reconstruction. (e) Grasp Transfer.

Fig. 2: *Multi-feature Implicit Model* (MIMO) and its applications. (a) MIMO takes as input an object point cloud $\mathbf{P}$ and a point coordinate $\mathbf{x}$ and outputs multiple spatial features of $\mathbf{x}$ relative to $\mathbf{P}$, including occupancy $\mathbf{\Phi}_{\text{occ}}$, signed distance $\mathbf{\Phi}_{\text{sdf}}$, extended space coverage feature (ESCF) $\mathbf{\Phi}_{\text{escf}}$ and closest distance direction (CDD) $\mathbf{\Phi}_{\text{cdd}}$. The concatenation of activation layers of the decoder for $\mathbf{\Phi}_{\text{escf}}$ and $\mathbf{\Phi}_{\text{cdd}}$ forms the point descriptor $\mathbf{z}$ of $\mathbf{x}$. (b) The CDD is represented as the inner product of two unit vectors $\mathbf{v}_p$ and $\mathbf{v}_d$. (c) The high-dimensional point descriptors of each reference object are reduced to a 3D space using Principal Component Analysis (PCA) representing the RGB channels of the color map. Each point of other categorical object instances is colorized according to the most similar point (smallest L1 distance in point descriptors) from the corresponding reference object. The MIMO can be used for (d) object shape reconstruction and (e) grasp pose transfer.

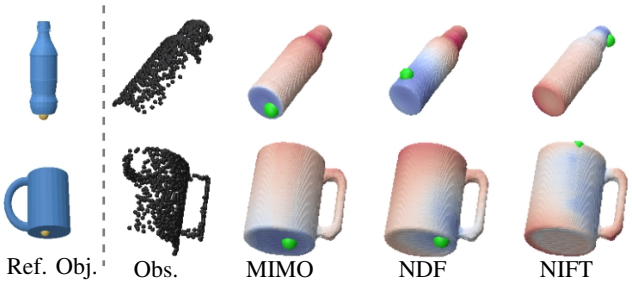


Fig. 3: Point correspondence and shape similarity measure using point descriptors from partially-observed point clouds (•). Given a point on a reference object, we colorize the novel object mesh based on the L1 distance of point descriptors to the reference point, where blue means more similar, and mark the most similar points (•).

III. MIMO for Manipulation

In this paper, we focus on learning task-oriented grasping and object rearrangement tasks from human demonstration videos. We first introduce the *Multi-feature Implicit Model* (MIMO) and its applications in Section III-A, and then propose a novel grasping framework in Section III-B to learn and generate task-oriented grasps.

A. Multi-feature Implicit Model

As shown in Fig. 2a, MIMO uses a shared PointNet [36] encoder $\varepsilon(\mathbf{P})$ embedding the geometric information of the point cloud $\mathbf{P}$ in a latent code, and a partly shared Multi-layer Perceptron (MLP) decoder with multiple branches, representing spatial relations of a point $\mathbf{x}$ relative to $\mathbf{P}$. The occupancy $\mathbf{\Phi}_{\text{occ}}$ [10] and signed distance $\mathbf{\Phi}_{\text{sdf}}$ [11] branches enable MIMO to reconstruct object shapes. Specifically, given the fully- or partially-observed point cloud of an object, we extract the object mesh from the trained occupancy branch using the Multi-resolution IsoSurface Extraction algorithm [10] (see Fig. 2d). We experience that, jointly training the signed distance branch yields more precise shape reconstruction compared with training the occupancy branch alone. Additionally, we introduce two novel feature branches, namely, 1) the extended SCF (ESCF) branch $\mathbf{\Phi}_{\text{escf}}$; and 2) the closest distance direction (CDD) branch $\mathbf{\Phi}_{\text{cdd}}$. In contrast to the SCF branch utilized in NIFT [8], where the power spectrum of each degree in the spherical harmonics expansion is considered, our ESCF branch is directly supervised by the coefficients of spherical harmonics expansion across all orders and degrees. This enables ESCF to capture finer geometric details. To further enhance the neural field's direction-awareness, we introduce CDD, defined as the inner product of unit vectors $\mathbf{v}_d$ and $\mathbf{v}_p$, where $\mathbf{v}_d$ points from a point $\mathbf{x}$ to the closest point on the object, and $\mathbf{v}_p$ follows a chosen principal direction, e. g., pointing upward when the object is positioned vertically (see Fig. 2b). Similarly to NDF, we concatenate the activation layers of the *partly-shared decoder* for $\mathbf{\Phi}_{\text{escf}}$ and $\mathbf{\Phi}_{\text{cdd}}$ as the point descriptor $\mathbf{z} = \kappa(\mathbf{x}|\mathbf{P})$, which forms a descriptor space to measure geometric similarity (see Fig. 2c). Trained with four branches, our descriptor space is more informative in distinguishing fine geometric details. In practice, we observed that the performance of the similarity measure drops when directly inferring $\mathbf{z}$ from the noisy partially-observed point cloud $\mathbf{P}$. To address this problem, we reconstruct the mesh, from which a point cloud $\mathbf{P}_r$ is sampled as input to MIMO to infer the point descriptor $\mathbf{z} = \kappa(\mathbf{x}|\mathbf{P}_r)$. As shown in Fig. 3, MIMO finds point correspondence between the reference object and a partially-observed categorical object precisely, while NDF yields an imprecise point correspondence and NIFT often fails to distinguish the up and down direction of the bottle or the mug. Further evaluation results are shown in Section IV-A. Since all the features can be automatically computed, no further human annotation is required to collect the training dataset. Next, we detail the loss functions for training MIMO.

1) Multi-task Loss Function: For training MIMO, having four distinct feature branches, we combine the loss functions of each branch through a weighted sum. However, manually

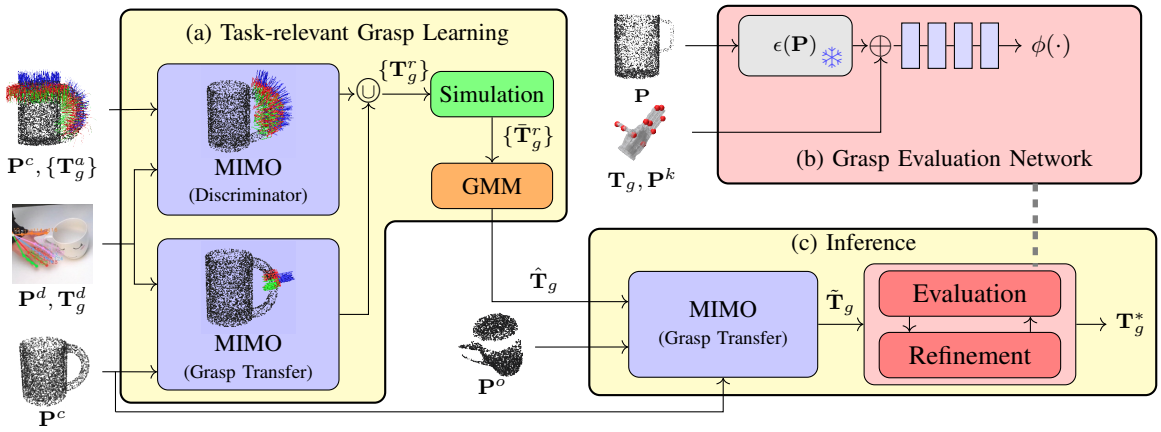


Fig. 4: Proposed MIMO-based Grasp Framework. (a) Given a human demonstration of a grasping scene, we obtain the object point cloud $\mathbf{P}^d$ and a grasp pose $\mathbf{T}_g^d$. We generate task-agnostic grasp poses $\{\mathbf{T}_g^a\}$ using a grasp generator [38], and use MIMO as a discriminator to select the task-relevant candidates $\{\mathbf{T}_g^r\}$ based on pose descriptor similarities between $\mathbf{T}_g^d$ and $\mathbf{T}_g^a$. Alternatively, we can directly transfer the demonstrated grasp pose $\mathbf{T}_g^d$ to the canonical point cloud $\mathbf{P}^c$ using MIMO. We then simulate the candidates $\{\mathbf{T}_g^r\}$ to find the successful grasp poses $\{\bar{\mathbf{T}}_g^r\}$ to train a GMM. (b) Given an object point cloud $\mathbf{P}$, a grasp pose $\mathbf{T}_g$ and a set of hand keypoints $\mathbf{P}^k$, the grasp evaluation network encodes $\mathbf{P}$ using the frozen encoder $\epsilon(\cdot)$ of MIMO and outputs the grasp success probability using MLP. (c) During inference, the sampled grasp pose $\hat{\mathbf{T}}_g$ relative to the canonical point cloud $\mathbf{P}^c$ is transferred to a partially-observed point cloud $\mathbf{P}^o$ using MIMO, and the transferred grasp pose $\tilde{\mathbf{T}}_g$ is evaluated and refined (if necessary) to obtain the optimal grasp pose $\mathbf{T}_g^*$.

adjusting the weights for these loss functions is challenging. To address this problem, we introduce homoscedastic uncertainty [37] for each branch, where the likelihood is defined as a Gaussian $p(\mathbf{y_i}|f_{\mathbf{W_i}}(\mathbf{x})) = \mathcal{N}(f_{\mathbf{W_i}}(\mathbf{x}), \sigma_i^2), i \in [1, 4]$ with the model output $f_{\mathbf{W_i}}(\mathbf{x})$ as the mean and the variance σ_i representing the uncertainty. The objective is to maximize the overall likelihood, or equivalently to minimize its negative log-likelihood, i. e., $\mathcal{L} = \sum_{i=1}^{4}(\frac{1}{2\sigma_i^2}\mathcal{L}_i(\mathbf{W_i}) + \log(\sigma_i))$, where $\mathcal{L}_i$ are binary cross entropy loss for occupancy, clamped L1 loss for signed distance, and L1 losses for ESCF and CDD, respectively. For better numerical stability, we set $s_i = \log(\sigma_i^2), i = \{1, 2, 3, 4\}$ following [37]. Thus, the total loss is reformulated as $\mathcal{L} = \sum_{i=1}^{4}(e^{-s_i}\mathcal{L}_i(\mathbf{W_i}) + s_i)$. During the training, we minimize the loss function with respect to weights of the model $\mathbf{W_i}$ and uncertainty s_i. In this way, uncertainty is automatically optimized without manual tuning.

2) Pose Descriptor: Similarly to [7], we adopt the Basis Point Set (BPS) [39] sampling strategy, and concatenate the point descriptors of a set of points around an object as their pose descriptor $\mathbf{Z}$. Specifically, given a set of points $\mathbf{X} \in \mathbb{R}^{N\times 3}$ sampled from a rigid object $\mathcal{O}_B$ in pose $\mathbf{T}$ around the point cloud $\mathbf{P}^A$ of object $\mathcal{O}_A$, we obtain pose descriptor of $\mathcal{O}_B$ using the trained MIMO of object category A, i. e., ${}^A\mathbf{Z}_B = \varphi(\mathbf{T}, \mathbf{X}|\mathbf{P}^A)$. It measures the similarity of the poses relative to $\mathcal{O}_A$, where similar poses have a small L1 distance between their pose descriptors. Speaking in terms of the example in Fig. 1, $\mathcal{O}_A$ would be an instance of the "mug" class, while $\mathcal{O}_B$ would be the hand and, therefore, ${}^A\mathbf{Z}_B$ associated to a grasp pose $\mathbf{T}$. Similarly to Section III-A, we reconstruct the mesh of $\mathcal{O}_A$, from which a point cloud $\mathbf{P}_r^A$ is sampled as input to MIMO to infer the pose descriptor ${}^A\mathbf{Z}_B = \varphi(\mathbf{T}, \mathbf{X}|\mathbf{P}_r^A)$.

3) Pose Transfer: Given a trained MIMO of object category A, a reference pose descriptor ${}^A\hat{\mathbf{Z}}_B$ and a pair of arbitrary object instances ($\bar{\mathcal{O}}_A$,$\bar{\mathcal{O}}_B$) from category A and B, we optimize the pose of $\bar{\mathcal{O}}_B$ relative to $\bar{\mathcal{O}}_A$ by $\mathbf{T}^* = \arg\min_{\mathbf{T}} \|\varphi(\mathbf{T}, \mathbf{X}|\bar{\mathbf{P}}_r^A) - {}^A\hat{\mathbf{Z}}_B\|_1$, where $\bar{\mathbf{P}}_r^A$ is the reconstructed point cloud of $\bar{\mathcal{O}}_A$. We adopt the same optimization procedure as in [7]. In a visual imitation learning (VIL) setup, the reference pose descriptor can be derived from human demonstration videos. Specifically, we find the closest point pair on $\mathcal{O}_A$ and $\mathcal{O}_B$ at the last timestep of the demonstration as keypoints. Similarly to [9], we then sample BPS around keypoints of $\mathcal{O}_A$ and $\mathcal{O}_B$ respectively, to compute the corresponding reference pose descriptors. which can be used to transfer $\bar{\mathcal{O}}_A$ and $\bar{\mathcal{O}}_B$ to align with $\mathcal{O}_A$ and $\mathcal{O}_B$, respectively, with the optimization steps described above. The rearrangement target pose of $\bar{\mathcal{O}}_B$ relative to $\bar{\mathcal{O}}_A$ can be derived from the optimized poses. Note that the keypoints and sampled BPS do not need to lie on the object. We refer interested readers to [9] for more details. In terms of grasping, where $\mathcal{O}_B$ is the human or robot hand and $\mathcal{O}_A$ is an arbitrary object to be grasped, the pose descriptors measure the grasp similarity, which can be used for transferring grasps to similar objects. Next, we introduce a novel grasp framework based on MIMO.

B. MIMO-based Grasp Framework

Leveraging MIMO's strengths in measuring pose similarities and transferring poses, we introduce a framework designed to learn task-specific grasping and object rearrangement from one or multiple human demonstrations. This framework can generate optimal grasp poses for new object instances based on partial observations, as shown in Fig. 4.

1) Human Observation: Given human demonstration videos consisting of sequences of RGB and depth images of a manipulation task, we estimate the hand poses in all frames using [40] and train a movement primitive [41] representing the hand motion. We then determine the grasping timestep t_g and detect grasp pose $\mathbf{T}_g \in SE(3)$ following [42]. The object being grasped is the source object $\mathcal{O}_s$, and the other object, which sets a reference frame for placing $\mathcal{O}_s$ at the last timestep t_T, is the target object $\mathcal{O}_t$. We obtain the segmented point clouds of both objects at t_g and t_T using Grounded SAM [43], [44].

2) Task-oriented Grasp Learning: As shown in Fig. 4 (a), we generate task-agnostic grasp candidates $\{\mathbf{T}_g^a\}$ using [38] on a canonical point cloud $\mathbf{P}^c$ for the class of the source object $\mathcal{O}_s$. We present two strategies to obtain task-relevant grasp candidates, i. e., (i) using MIMO as a discriminator for pose similarity to find the most similar grasps in $\{\mathbf{T}_g^a\}$ to $\mathbf{T}_g^d$ (see Section III-A.2); or (ii) using MIMO to directly transfer the demonstrated grasp $\mathbf{T}_g^d$ relative to $\mathbf{P}^d$ to a set of candidate grasps relative to canonical space (see Section III-A.3). We fuse the task-relevant grasp candidates $\{\mathbf{T}_g^r\}$ from the two strategies and simulate them with a humanoid hand in Issac Gym [45]. Specifically, the grasp is successful if the object is picked up and does not drop after a random shaking action. We then simulate the object rearrangement given the successful grasps and determine the set of task-relevant grasps if the tasks are accomplished without failure (see Section IV for a definition of possible tasks). The successful and task-relevant grasps $\{\bar{\mathbf{T}}_g^r\}$ in canonical space are used to train a GMM on a Riemannian manifold (i. e., $\mathbb{R}^3 \times \mathcal{S}^3$), which can be used to generate task-oriented grasps.

3) Grasp Evaluation: The sampled task-relevant grasps from the GMM are not guaranteed to be successful. To address this problem, we propose a *task-agnostic* grasp evaluation network to compute the success probability of a grasp pose $\mathbf{T}_g$ relative to an arbitrary point cloud $\mathbf{P}$ (see Fig. 4). We first encode $\mathbf{P}$ using the frozen encoder of MIMO, i. e., $\mathbf{c} = \varepsilon(\mathbf{P})$. We then use a MLP decoder conditioned on $\mathbf{c}$ to predict the success probability given a set of keypoints $\mathbf{P}^k$ on the humanoid hand representing its pose, i. e., $\phi(\mathbf{T}_g, \mathbf{P}^k | \varepsilon(\mathbf{P})) \in [0, 1]$. We train this model using a binary cross-entropy loss on a dataset fusing the task-agnostic grasp candidates for all objects in all tasks, along with their binary labels indicating successful grasps obtained in Section III-B.2.

4) Inference: During inference, we sample grasp poses $\hat{\mathbf{T}}_g$ from the trained GMM relative to the canonical point cloud $\mathbf{P}^c$, which are then transferred to a partially-observed point cloud $\mathbf{P}^o$ of a novel categorical instance following Section III-A.3. We compute the success probability of the transferred grasps $\hat{\mathbf{T}}_g$ using the trained task-agnostic grasp evaluation network. If the grasp success probability is lower than a certain threshold ξ, we refine the grasp pose by maximizing the grasping success likelihood using the grasp evaluation network from Section III-B.3, i. e., $\Delta\mathbf{T}_g^* = \arg\max_{\Delta\mathbf{T}_g} \phi(\Delta\mathbf{T}_g\mathbf{T}_g, \mathbf{P}^k | \varepsilon(\mathbf{P}))$, and finally obtain the optimal grasp pose $\mathbf{T}_g^* = \Delta\mathbf{T}_g^*\mathbf{T}_g$.

IV. Evaluation

We evaluate the proposed MIMO and grasping framework in different manipulation tasks and compare with NDF [7], R-NDF [9], and NIFT [8]. More details, evaluation videos, and source code are available at https://sites.google.com/view/mimo4.

A. Evaluation of MIMO in Simulation

We first evaluate the performance of MIMO against different approaches from the state of the art. To show the effectiveness of the novel ESCF and CDD features in MIMO (denoted MIMO4), we provide additional evaluation results of a variant of MIMO (denoted MIMO3) with three branches in the decoder to predict occupancy, signed distance, and SCF separately.

1) Generation of Training Data: Training MIMO can be done without manual annotation of the training data. NDF and NIFT each provide their own datasets that could be used for training. However, we observed two issues in these datasets, namely (i) the bottom of the bottle's meshes from NDF is hollowed out, which influences the shape reconstruction quality; (ii) the mesh scaling is non-uniform, leading to wrong labels for SCF and signed distance. Therefore, we generate a new dataset made from watertight meshes from the ShapeNet dataset [46] using [47] with rendered point clouds for each mesh. The remainder of the data generation and training of the models is similar to the procedure used for NIFT. We train NIFT and our model using the new dataset on a single NVIDIA A100 GPU, and use the pre-trained weights of NDF and R-NDF provided by the authors.

2) Setup and Metrics: We consider three settings, namely (S1) 10 demonstrations and four viewpoints, where the point cloud is fused from 4 depth cameras at 4 corners of the table; (S2) a single demonstration and four viewpoints, with the same camera positions as before; and (S3) a single demonstration and single viewpoint, in which the mug handle and bottle opening are visible. We use BPS for all models in the evaluation tasks. To evaluate SE(3)-equivariance of the trained neural fields, we distinguish between upright (U) and arbitrary (A) initial object poses, where the objects are positioned upright on the table for U while the objects are arbitrarily positioned in the air for A. For MIMO4 and MIMO3, we reconstruct object shapes from partial observations and transfer poses as discussed in Section III-A.3. The overall task is successful if the object is grasped without dropping (grasp success) and the bowl/bottle stands upright on a shelf, or the mug is hung on the rack without penetration at the optimized target pose (placement success).

3) Comparison with NDF: We use the simulation environment and evaluation proposed from NDF, including 3 pick-and-place tasks: (T1) picking a mug by the rim and placing it on the rack by the handle; (T2) picking a bowl and placing it on the shelf; and (T3) picking a bottle from the side and placing it on the shelf. We conduct 100 trials for each task under the two settings (S1) and (S3), and upright and arbitrary object poses respectively.

TABLE I: Unseen object pick-and-place success rate with setting (S3) (single viewpoint, single demonstration).

		Mug (T1)			Bowl (T2)			Bottle (T3)			Mean		
		Grasp	Place	Overall	Grasp	Place	Overall	Grasp	Place	Overall	Grasp	Place	Overall
Upr. Pose U	NDF	0.95	0.73	0.72	0.89	0.93	0.84	0.90	0.69	0.65	0.91	0.78	0.74
	R-NDF	0.90	0.77	0.69	0.90	**1.00**	0.90	0.53	**0.97**	0.51	0.78	0.91	0.70
	NIFT	0.99	0.92	0.92	0.98	**1.00**	0.98	0.96	0.94	0.90	0.98	0.95	0.93
	MIMO3	**1.00**	0.92	0.92	0.99	**1.00**	**0.99**	0.92	0.93	0.91	0.97	0.95	0.94
	MIMO4	**1.00**	**0.98**	**0.98**	**1.00**	0.99	**0.99**	**0.97**	**0.97**	**0.95**	**0.99**	**0.98**	**0.97**
Arb. Pose A	NDF	0.53	0.58	0.34	0.76	0.80	0.64	0.42	0.91	0.40	0.57	0.76	0.46
	R-NDF	0.50	0.70	0.35	0.78	0.97	0.77	0.12	0.90	0.09	0.47	0.86	0.40
	NIFT	0.46	0.90	0.42	0.96	0.96	0.94	0.38	0.93	0.37	0.60	0.93	0.58
	MIMO3	0.86	0.94	0.80	0.94	**0.99**	0.94	0.77	0.87	0.71	0.86	0.93	0.82
	MIMO4	**0.92**	**0.97**	**0.90**	**0.98**	0.97	**0.95**	**0.95**	**0.97**	**0.93**	**0.95**	**0.97**	**0.93**

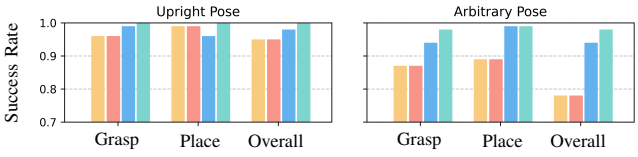


Fig. 5: Success rate of the pick-and-place tasks (T1)-(T3) with unseen objects under setting (S1) for models NDF , NIFT , MIMO3 , and MIMO4 , respectively.

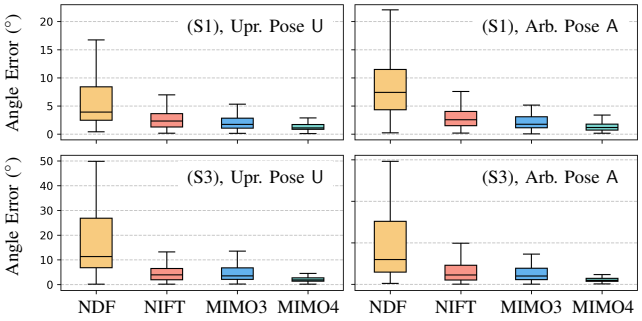


Fig. 6: Angle error of bowls and bottles. Colors as Fig. 5.

As shown in Fig. 5, all approaches achieve high success rates of tasks (T1)-(T3) for the setting (S1). We observe that MIMO4 achieves the best result in all cases, with MIMO3 ranked slightly below. It is important to note that the overall success rates of MIMO4 only drop by 2% in arbitrary pose compared to upright pose, which is less than other approaches, showcasing a better SE(3)-equivariance property of our neural descriptor field. In contrast, as shown in Table I, MIMO4 significantly outperforms others in setting (S3), especially in the case of arbitrary object poses in tasks (T1) and (T2). We highlight the best success rates of each (sub-)task. MIMO3, NIFT and R-NDF only perform slightly or equally well than MIMO4 in the placing phase of (T2), where bowls are involved. The reason is that the partially-observed point cloud of bowls with large opening has already covered a large portion of the object and it is much easier to distinguish the up and down direction compared to the mugs and bottles used in (T1) and (T2). As discussed in Section III-A and Fig. 3, NDF and NIFT often fail to distinguish the top and bottom of the bottle and mug handle. Inaccurate correspondences can cause objects to be transformed into wrong poses, leading to low success rates in (T1) and (T2). In contrast, our descriptor field is more informative, achieving more accurate pose transfer and thus higher success rates. We showcase the pose accuracy in Fig. 6 by computing the angle error between the object's upright direction and the gravity direction at the target pose for bowls and bottles in (T2) and (T3). A smaller angle indicates a more precise placement pose. We observe that our MIMO4 has the smallest average angle error and smallest variance across all tasks, further verifying the superiority of our neural descriptor.

TABLE II: Success rates of unseen object rearrangement. U and A stand for upright and arbitrary poses, respectively.

		(T4)		(T5)		(T6)		Mean	
	Models	U	A	U	A	U	A	U	A
(S1)	R-NDF	0.71	0.55	0.75	0.75	0.80	0.54	0.75	0.61
	MIMO3	**0.91**	**0.87**	**0.92**	**0.91**	0.84	0.85	**0.89**	0.88
	MIMO4	0.88	0.85	0.91	0.89	**0.87**	**0.93**	**0.89**	**0.89**
(S2)	R-NDF	0.56	0.53	0.64	0.61	0.12	0.18	0.44	0.44
	MIMO3	0.89	0.89	**0.90**	**0.88**	0.85	0.87	0.88	0.88
	MIMO4	**0.92**	**0.92**	**0.90**	0.87	**0.91**	**0.93**	**0.91**	**0.92**
(S3)	R-NDF	0.29	0.21	0.10	0.13	0.16	0.07	0.18	0.14
	MIMO3	0.85	0.85	0.88	0.87	0.72	0.70	0.82	0.81
	MIMO4	**0.89**	**0.86**	**0.90**	**0.88**	**0.90**	**0.83**	**0.90**	**0.86**

4) Comparison with R-NDF: We adopt the simulation environments from R-NDF [9] with three tasks, namely: (T4) hanging a mug on the hook of a rack; (T5) placing a bowl on a mug; and (T6) and placing a bottle in a container. All three settings are considered, namely (S1), (S2) and (S3). In contrast to the experiments in Section IV-A.3, we focus only on the target configurations of the object and neglect the grasp procedure for this evaluation. The task is successful if the source object is placed on the target object without falling or exerting a large interaction force. We conduct 100 trials for each task and compute the success rates. As shown in Table II, MIMO4 and MIMO3 perform equally well in setting (S1) with a success rate of about 89%. In both settings (S2) and (S3), MIMO4 significantly outperforms R-NDF by about 48% and 70%, respectively. Therefore, we do not need the extra alignment and refinement steps as in NIFT [9]. Note that MIMO3's performance drops in (S2) and further in (S3), showcasing the effectiveness of the novel ESCF and CDD feature in the partly shared decoder of MIMO.

B. Evaluation of the Grasping Framework

To evaluate the performance of MIMO in the context of our grasping framework presented in Section III-B, we

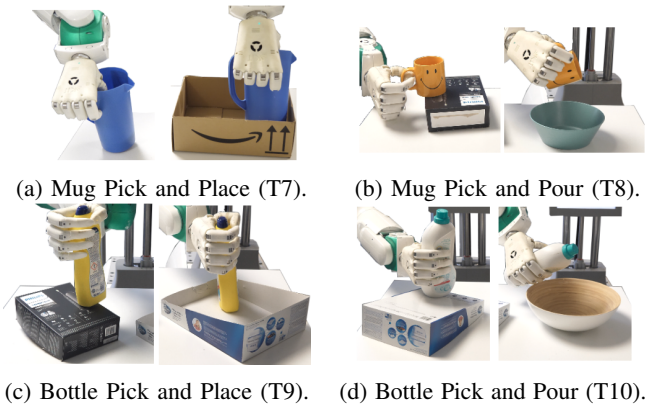
TABLE III: The success rates of unseen object grasping (G) and rearrangement (R).

	(T7)		(T8)		(T9)		(T10)		Mean	
Models	G	R	G	R	G	R	G	R	G	R
NIFT	0.80	0.62	0.92	0.80	0.86	0.08	0.92	0.68	0.88	0.55
MIMO4	**0.94**	**0.88**	**0.96**	**0.94**	**0.90**	**0.80**	**0.98**	**0.88**	**0.95**	**0.88**

performed multiple experiments in simulation using Isaac Gym [45] and on the humanoid robots ARMAR-6 [48] and ARMAR-DE in real-world manipulation tasks. We define four tasks, namely: (T7) grasp a mug at its rim and place it upright in a container; (T8) grasp a mug at its handle and pour into a bowl; (T9) grasp a bottle at its neck and place it upright in a container; and (T10) grasp a bottle at its body and pour it into a bowl. Object poses are randomly initialized, with mugs positioned to ensure handle visibility. We use MIMO4 to reconstruct object shapes from the partially-observed object point clouds. The grasp poses are sampled from the GMM, transferred to the observed objects, and evaluated by the grasp evaluator (see Section III-B.3). If the estimated success probability drops below 0.9, the grasp pose is optimized with a learning rate of 10^{-3} (see III-B.4). We then optimize the target pose for rearrangement using MIMO4 and execute the grasp and rearrangement action.

1) Evaluation in Simulation: We simulate a humanoid hand in Issac Gym equipped with a depth camera positioned in front of a table. We use NIFT with BPS as a baseline approach without the grasp evaluation and refinement. Note that tasks (T7)-(T10) are successful if both grasping and rearrangement are successful. We execute each task for 50 trials under setting (S3). As shown in Table III, MIMO4 outperforms NIFT in all tasks, especially in task (T9) by about 72%. NIFT cannot differentiate between the top and bottom of the bottle and, therefore, fails to place the bottle in the container. In contrast, MIMO4 achieves higher success rates, benefiting from the reconstructed shape and the powerful descriptor space. In addition, MIMO4 achieves an average success rate of 95% for grasping, including difficult side grasps at the mug handle, which demonstrates the effectiveness of our grasp evaluator.

2) Evaluation in the Real World: Similarly to Section IV-B.1, we replicate tasks (T7)-(T10) using the same GMM and MIMO4 with two humanoid robots: ARMAR-DE and ARMAR-6. We use an Azure Kinect camera mounted on the robot head to obtain RGB and depth images and extract object point clouds as explained in Section III-B.1. For the experiments on ARMAR-DE, the grasp pose was validated and executed using the mobile manipulation framework [49]. On ARMAR-6, we use a task-space impedance controller to execute the motions generated by the learned movement primitives (see Section III-B.1), where the target poses are the corresponding grasp pose in the grasp phase and the object rearrangement pose in the placement or pouring phase. We show qualitative results in Fig. 7 and in the accompanying video, showcasing the efficacy of our approach in one-shot imitation learning of manipulation tasks.



(a) Mug Pick and Place (T7). (b) Mug Pick and Pour (T8).

(c) Bottle Pick and Place (T9). (d) Bottle Pick and Pour (T10).

Fig. 7: Real-world experiments on ARMAR-DE.

V. Conclusion

We propose *Multi-feature Implicit Model* (MIMO), a novel implicit neural field that provides informative and SE(3)-equivariant point and pose descriptors for shape similarity measure. Trained on multiple spatial features, MIMO facilitates finer correspondence detection and more accurate pose transfer compared to state-of-the-art approaches. MIMO also allows for shape reconstruction to account for partial observations. Based on MIMO, we propose a task-oriented grasping and object rearrangement framework with a novel evaluation and refinement procedure to further increase success rates. Our approach outperforms others in the one- and few-shot visual imitation learning of pick-and-rearrangement tasks. In future works, we will investigate local neural descriptors and inter-category generalization of manipulation skills.

References

[1] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic, "Affordance Detection for Task-specific Grasping Using Deep Learning," in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2017, pp. 91–98.

[2] R. Detry, J. Papon, and L. Matthies, "Task-oriented Grasping with Semantic and Geometric Scene Understanding," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017, pp. 3266–3273.

[3] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based Affordances Detection with Convolutional Neural Networks and Dense Conditional Random Fields," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017, pp. 5908–5915.

[4] J. Jin and M. Jagersand, "Generalizable Task Representation Learning from Human Demonstration Videos: A Geometric Approach," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 2504–2510.

[5] J. Gao, Z. Tao, N. Jaquier, and T. Asfour, "K-VIL: Keypoints-based Visual Imitation Learning," *IEEE Trans. on Robotics*, vol. 39, no. 5, pp. 3888–3908, 2023.

[6] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural Fields in Visual Computing and Beyond," in *Computer Graphics Forum*, vol. 41, no. 2, 2022, pp. 641–676.

[7] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 6394–6400.

[8] Z. Huang, J. Xu, S. Dai, K. Xu, H. Zhang, H. Huang, and R. Hu, "NIFT: Neural Interaction Field and Template for Object Manipulation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2023, pp. 1875–1881.

[9] A. Simeonov, Y. Du, Y.-C. Lin, A. R. Garcia, L. P. Kaelbling, T. Lozano-Pérez, and P. Agrawal, "SE(3)-Equivariant Relational Rearrangement with Neural Descriptor Fields," in *Conference on Robot Learning (CoRL)*, 2023, pp. 835–846.

[10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy Networks: Learning 3D Reconstruction in Function Space," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4460–4470.

[11] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 165–174.

[12] S. Pfrommer, M. Halm, and M. Posa, "Contactnets: Learning Discontinuous Contact Dynamics with Smooth, Implicit Representations," in *Conference on Robot Learning (CoRL)*, 2021, pp. 2279–2291.

[13] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang, "Grasping Field: Learning Implicit Representations for Human Grasps," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 333–344.

[14] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation," in *Conference on Robot Learning (CoRL)*, 2018, pp. 373–385.

[15] C.-Y. Chai, K.-F. Hsu, and S.-L. Tsao, "Multi-step Pick-and-place Tasks using Object-centric Dense Correspondences," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 4004–4011.

[16] S. Yang, W. Zhang, R. Song, J. Cheng, and Y. Li, "Learning Multi-object Dense Descriptor for Autonomous Goal-conditioned Grasping," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4109–4116, 2021.

[17] X. Zhao, R. Hu, P. Guerrero, N. Mitra, and T. Komura, "Relationship Templates for Creating Scene Variations," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–13, 2016.

[18] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An Affordance keypoint Detection Network for Robot Manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.

[19] P. Ardón, E. Pairet, R. P. Petrick, S. Ramamoorthy, and K. S. Lohan, "Learning Grasp Affordance Reasoning Through Semantic Relations," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4571–4578, 2019.

[20] R. Monica and J. Aleotti, "Point Cloud Projective Analysis for Part-Based Grasp Planning," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4695–4702, 2020.

[21] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations," *Robotics: Science and Systems (R:SS)*, 2021.

[22] W. Chen, H. Liang, Z. Chen, F. Sun, and J. Zhang, "Learning 6-DoF Task-oriented Grasp Detection via Implicit Estimation and Visual Affordance," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2022, pp. 762–769.

[23] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning Task-oriented Grasping for Tool Manipulation from Simulated Self-supervision," *Intl. Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

[24] B. Wen, W. Lian, K. Bekris, and S. Schaal, "CaTGrasp: Learning Category-Level Task-Relevant Grasping in Clutter from Simulation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 6401–6408.

[25] V. Holomjova, A. J. Starkey, B. Yun, and P. Meißner, "One-shot Learning for Task-oriented Grasping," *IEEE Robotics and Automation Letters*, 2023.

[26] D. Hadjivelichkov, S. Zwane, L. Agapito, M. P. Deisenroth, and D. Kanoulas, "One-Shot Transfer of Affordance Regions? AffCorrs!" in *Conference on Robot Learning (CoRL)*, 2023, pp. 550–560.

[27] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language Embedded Radiance Fields for Zero-shot Task-oriented Grasping," in *Conference on Robot Learning (CoRL)*, 2023, pp. 178–200.

[28] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "LERF: Language Embedded Radiance Fields," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 19 729–19 739.

[29] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "KPAM: KeyPoint Affordances for Category-Level Robotic Manipulation," in *Intl. Symp. on Robotics Research*, 2019, pp. 132–157.

[30] W. Gao and R. Tedrake, "KPAM-SC: Generalizable Manipulation Planning using Keypoint Affordance and Shape Completion," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 6527–6533.

[31] W. Gao and R. Tedrake, "KPAM 2.0: Feedback Control for Category-Level Robotic Manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2962–2969, 2021.

[32] D. Rodriguez, C. Cogswell, S. Koo, and S. Behnke, "Transferring Grasping Skills to Novel Instances by Latent Space Non-rigid Registration," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 4229–4236.

[33] D. Rodriguez and S. Behnke, "Transferring Category-based Functional Grasping Skills by Latent Space Non-rigid Registration," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2662–2669, 2018.

[34] O. Biza, S. Thompson, K. R. Pagidi, A. Kumar, E. van der Pol, R. Walters, T. Kipf, J.-W. van de Meent, L. L. Wong, and R. Platt, "One-shot Imitation Learning via Interaction Warping," *arXiv preprint arXiv:2306.12392*, 2023.

[35] D. Hidalgo-Carvajal, H. Chen, G. C. Bettelani, J. Jung, M. Zavaglia, L. Busse, A. Naceri, S. Leutenegger, and S. Haddadin, "Anthropomorphic Grasping with Neural Object Shape Completion," *IEEE Robotics and Automation Letters*, 2023.

[36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.

[37] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[38] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes," *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021.

[39] S. Prokudin, C. Lassner, and J. Romero, "Efficient Learning on Point Clouds with Basis Point Sets," in *Intl. Conf. on Computer Vision (ICCV)*, 2019, pp. 4332–4341.

[40] K. Lin, L. Wang, and Z. Liu, "Mesh Graphormer," in *Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 12 919–12 928.

[41] Y. Zhou, J. Gao, and T. Asfour, "Learning Via-Point Movement Primitives with Inter- and Extrapolation Capabilities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4301–4308.

[42] J. Gao, X. Jin, F. Krebs, N. Jaquier, and T. Asfour, "Bi-KVIL: Keypoints-based Visual Imitation Learning of Bimanual Manipulation Tasks," *arXiv:2303.07399*, 2024.

[43] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-set Object Detection," *arXiv:2303.05499*, 2023.

[44] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[45] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac Gym: High Performance GPU Based Physics Simulation for Robot Learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[46] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "ShapeNet: An Information-Rich 3D Model Repository," *arXiv preprint arXiv:1512.03012*, 2015.

[47] D. Stutz and A. Geiger, "Learning 3D Shape Completion from Laser Scan Data with Weak Supervision," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1955–1964.

[48] T. Asfour, M. Wächter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, "ARMAR-6: A High-Performance Humanoid for Human-Robot Collaboration in Real World Scenarios," *IEEE Robotics and Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.

[49] C. Pohl, F. Reister, F. Peller-Konrad, and T. Asfour, "MAkEable: Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills," *arXiv preprint arXiv:2401.16899*, 2024.