# D$^3$Fields: Dynamic 3D Descriptor Fields for Zero-Shot Generalizable Robotic Manipulation

Yixuan Wang[1*], Zhuoran Li[2,3*], Mingtong Zhang[1], Katherine Driggs-Campbell[1],
Jiajun Wu[2], Li Fei-Fei[2], Yunzhu Li[1,2]

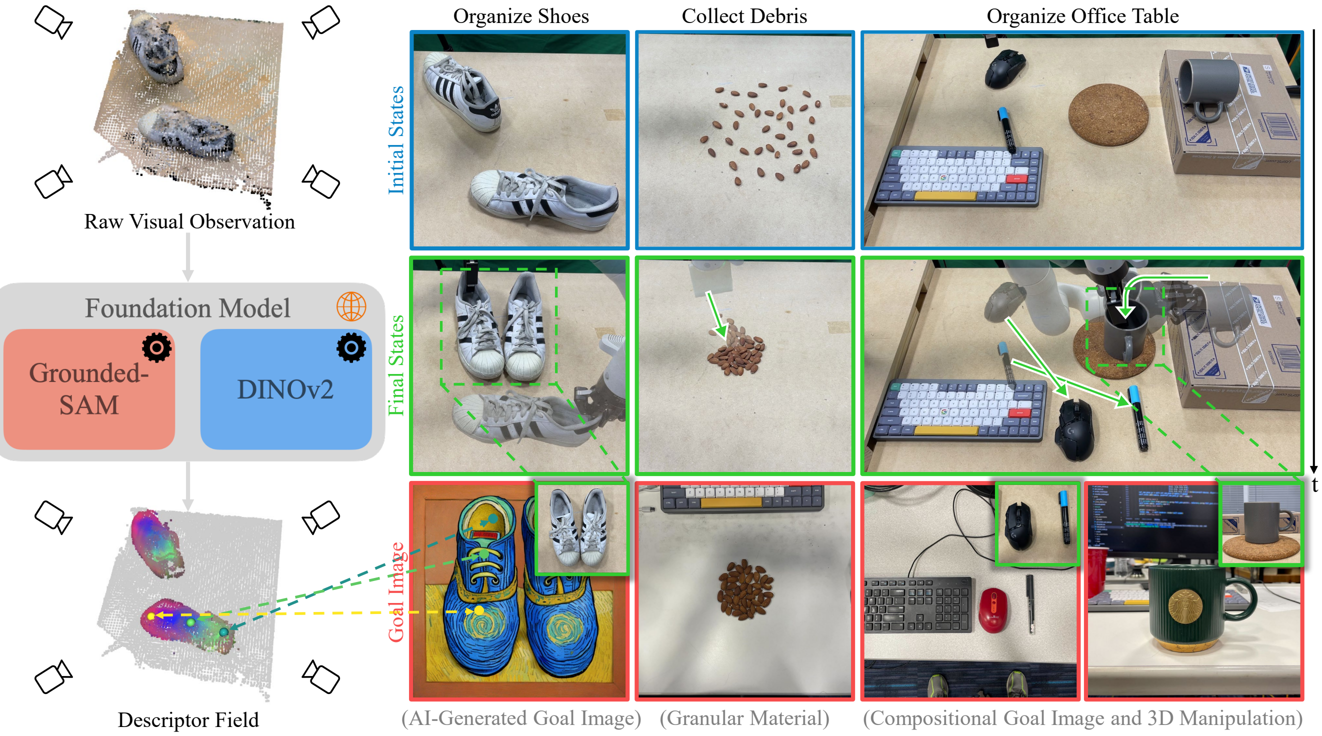


Fig. 1: **D$^3$Fields Representation and Application to Various Manipulation Tasks.** D$^3$Fields take in multi-view RGBD images and encode semantic features and instance masks using foundational models. The gray and colored points in the bottom left visualize background and semantic features mapped to RGB space using Principal Component Analysis (PCA), demonstrating consistency across instances. We use our representation for diverse tasks in a zero-shot manner. These tasks are defined by 2D goal images with diverse instances and styles. We address pick-and-place tasks such as shoe organization and tasks requiring dynamic modeling like collecting debris. We also demonstrate in the office table organization that our framework can accomplish 3D manipulation and compositional task specification.

***Abstract*— Scene representation has been a crucial design choice in robotic manipulation systems. An ideal representation should be 3D, dynamic, and semantic to meet the demands of diverse manipulation tasks. However, previous works often lack all three properties simultaneously. In this work, we introduce D$^3$Fields — dynamic 3D descriptor fields. These fields capture the dynamics of the underlying 3D environment and encode both semantic features and instance masks. Specifically, we project arbitrary 3D points in the workspace onto multi-view 2D visual observations and interpolate features derived from foundational models. The resulting fused descriptor fields allow for flexible goal specifications using 2D images with varied contexts, styles, and instances. To evaluate the effectiveness of these descriptor fields, we apply our representation to a wide range of robotic manipulation tasks in a zero-shot manner. Through extensive evaluation in both real-world scenarios and simulations, we demonstrate that D$^3$Fields are both generalizable and effective for zero-shot robotic manipulation tasks. In quantitative comparisons with state-of-the-art dense descriptors, such as Dense Object Nets and DINO, D$^3$Fields exhibit significantly better generalization abilities and manipulation accuracy. Project Page: https://robopil.github.io/d3fields/***

## I. INTRODUCTION

The choice of scene representation is critical in robotic systems. An ideal representation should be simultaneously 3D, dynamic, and semantic to meet the needs of various robotic manipulation tasks in our daily lives. However, previous research into scene representations in robotics often does not encompass all three properties. Some representations exist in 3D space [1–4], yet they overlook semantic information. Others focus on dynamic modeling [5–8], but only consider 2D data. Some other works are limited by only

*Denotes equal contribution. https://robopil.github.io/d3fields/

[1]University of Illinois Urbana-Champaign [2]Stanford Univeristy [3]National University of Singapore

considering semantic information such as object instance and category [9–13].

In this work, we aim to satisfy all three criteria by introducing D$^3$Fields, unified descriptor fields that are 3D, dynamic, and semantic. D$^3$Fields take in arbitrary points in the 3D world coordinate frame and output both geometric and semantic information related to these points. This includes the instance mask, dense semantic features, and the signed distance to the object surface. Notably, deriving these descriptor fields requires no training and is conducted in a zero-shot manner using large foundational vision models and vision-language models (VLMs). Specifically, we first use Grounding-DINO [14], Segment Anything (SAM) [15], XMem [16], and DINOv2 [17] to extract information from multi-view 2D RGB images. We then project the 3D points back to each camera, interpolate to compute representations from each view, and fuse these data to derive the descriptors for the associated 3D points, as shown in Fig. 1 (left). By leveraging the dense semantic feature and instance mask of our representation, we can robustly track 3D points of the target object instance and train dynamics models. These learned dynamics models can then be incorporated into a Model-Predictive Control (MPC) framework to plan for manipulation tasks.

Notably, the derived representations allow for goal specification using 2D images sourced from the Internet, phones, or those generated by AI models. Such goal images have been challenging to manage with previous methods, because they contain varied styles, contexts, and object instances different from the robot's workspace. Our proposed D$^3$Fields can establish dense correspondences between the robot workspace and the target configurations. These correspondences give us the task objective, enabling us to plan the robot's actions with the learned dynamics model within the MPC framework. This task execution process does not require any further training, offering a flexible and convenient interface for humans to instruct robots.

We evaluate our method across a wide range of household robotic manipulation tasks in a zero-shot manner. These tasks include organizing shoes, collecting debris, and organizing office desks, as shown in Fig. 1 (right). Furthermore, we offer detailed quantitative comparisons between our method and other state-of-the-art dense descriptor techniques. Our results indicate that our approach significantly outperforms in terms of generalizability and manipulation accuracy.

To summarize our contributions: (1) We introduce a novel representation, D$^3$Fields, that is **3D**, **dynamic**, and **semantic**. (2) We present a novel and flexible goal specification method using 2D images that incorporate a range of styles, contexts, and instances. (3) Our proposed robotic manipulation framework supports zero-shot generalizable manipulation applicable to a broad spectrum of household tasks.

## II. Related Works

### A. Foundation Models for Robotics

Foundation models generally refer to those trained on broad data, often using self-supervision at scale, which can then be adapted (e.g., fine-tuned) to various downstream tasks. Large Language Models (LLMs) have showcased promising reasoning abilities for language. Robotics researchers have recently released a series of works that leverage LLMs, including SayCan [18] and Inner Monologue [19], to directly generate robot plans. Some later works have used LLMs as a code generator: Code as Policies [20] uses 2D object detectors as the perception API, whereas VoxPoser [21] creates a 3D value map. Yet, their perception modules fall short in modeling the precise geometry and dynamics of objects. Our D$^3$Fields aim to address this by focusing on detailed 3D geometry and dynamics.

Meanwhile, foundational vision models, such as SAM [15] and DINOv2 [17], have demonstrated impressive zero-shot generalization capabilities across various vision tasks. However, their focus is primarily on 2D vision tasks. Grounding these models in a dynamic 3D environment remains a challenge. The recent GROOT project showcases how to construct 3D object-centric representations using foundational models and exhibits notable few-shot generalization capabilities [22]. Still, GROOT does not emphasize learning about object dynamics or achieving zero-shot generalizable robotic manipulation.

### B. Representation for Visual Robotic Manipulation

Scene representation has been a pivotal component in robotic manipulation systems. Some early work relies on 2D representations, such as bounding boxes [23, 24]. Many recent methods construct particle representations of the environment and employ learned dynamics to capture the system's underlying structure [3, 7, 8, 25–29]. They demonstrate impressive results in unstructured environments and with non-rigid objects. However, they are not semantic, which can hinder their ability to generalize to new tasks and scenarios. Some research opts for a fixed-dimension latent vector derived from high-dimensional sensory inputs as the representation [2, 5, 6, 30–36], but such a representation does not scale well to complex manipulation tasks that require high precision and explicit scene structures. Other approaches use 6 DoF object poses as their representation [9, 10, 37, 38], though focusing primarily on grasping tasks instead of more dynamic ones. In this work, we aim to address these issues by introducing D$^3$Fields, a representation that models dynamic 3D environments at varying semantic levels.

### C. Neural Fields for Robotic Manipulation

Researchers have presented a variety of works using neural fields as a representation for robotic manipulation [39–41, 41–52]. Among them, Neural Descriptor Fields are the most relevant to ours [42]. They build neural feature fields that generalize to different instances with several demonstrations; but they focus on learning geometric, not semantic features, which hinders cross-category generalization.

Recently, a series of works distilled neural feature fields using foundation models such as CLIP and DINO for supervision [53, 54]. LeRF distills neural feature fields to handle open-vocabulary 3D queries and develops task-oriented

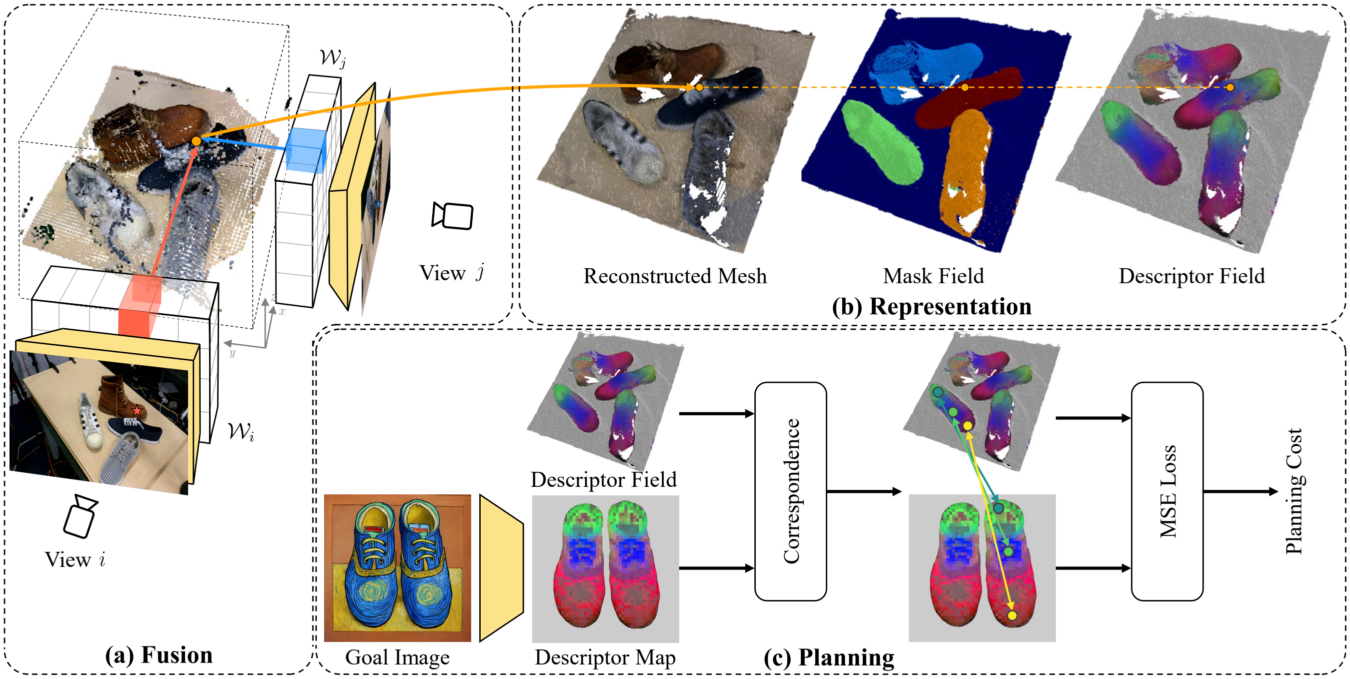


Fig. 2: **Overview of the proposed framework.** (a) The fusion process fuses RGBD observations from multiple views. Each view is processed by foundation models to obtain the feature volume $\mathcal{W}$. Arbitrary 3D points are processed through projection and interpolation. (b) After fusing information from multiple views, we obtain an implicit distance function to reconstruct the mesh form. We also have instance masks and semantic features for evaluated 3D points, as shown by the mask field and descriptor field in the top right subfigure. (c) Given a 2D goal image, we use foundation models to extract the descriptor map. Then we correspond 3D features to 2D features and define the planning cost based on the correspondence.

grasping based on it [55, 56]. Shen et al. [57] use a similar distilled feature field for the grasping task. Both methods require dense camera views to train the neural field. GNFactor addresses this by introducing a voxel encoder [58]. However, distilling foundation models to create neural feature fields has drawbacks: (1) They often require dense camera views for a quality field. (2) Distilled neural fields need retraining for new scenes, limiting their generalization and making them ineffective for dynamic scenes. In contrast, our D$^3$Fields do not need extra training for new scenes and can work with sparse views and dynamic settings.

## III. Method

In this section, we introduce the problem formulation in Section III-A and define camera transformation and projection notations in Section III-B. The construction of D$^3$Fields is detailed in Section III-C. Section III-D discusses tracking keypoints and learning dynamics, while Section IV-C showcases how our representation enables zero-shot generalizable manipulation skills.

### A. Problem Formulation

Given a 2D goal image $\mathcal{I}$, we denote the corresponding scene representation as $\boldsymbol{s}_{\text{goal}}$. Our goal is to find the action sequence $\{a^t\}$ to minimize the task objective:

$$\begin{aligned} \min_{\{a_t\}} \quad & c(\boldsymbol{s}^T, \boldsymbol{s}_{\text{goal}}), \\ \text{s.t.} \quad & \boldsymbol{s}^t = g(\boldsymbol{o}^t), \quad \boldsymbol{s}^{t+1} = f(\boldsymbol{s}^t, a^t), \end{aligned} \tag{1}$$

where $c(\cdot,\cdot)$ is the cost function measuring the distance between the terminal representation $\boldsymbol{s}^T$ and the goal representation $\boldsymbol{s}_{\text{goal}}$. Representation extraction function $g(\cdot)$ takes in the current multi-view RGBD observations $\boldsymbol{o}^t$ and outputs the current representation $\boldsymbol{s}^t$. $f(\cdot,\cdot)$ is the dynamics function that predicts the future representation $\boldsymbol{s}^{t+1}$, conditioned on the current representation $\boldsymbol{s}^t$ and action $a^t$. The optimization aims to find the action sequence $\{a_t\}$ that minimizes the cost function $c(\boldsymbol{s}^T, \boldsymbol{s}_{\text{goal}})$.

### B. Notation: Camera Transformation and Projection

We assume all cameras' intrinsic parameters $\mathbf{K}$ and extrinsic parameters $\mathbf{T}$ are known. The camera $i$ extrinsic parameters are defined as follows.

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ 0^T & 1 \end{bmatrix} \in \mathbb{SE}(3), \tag{2}$$

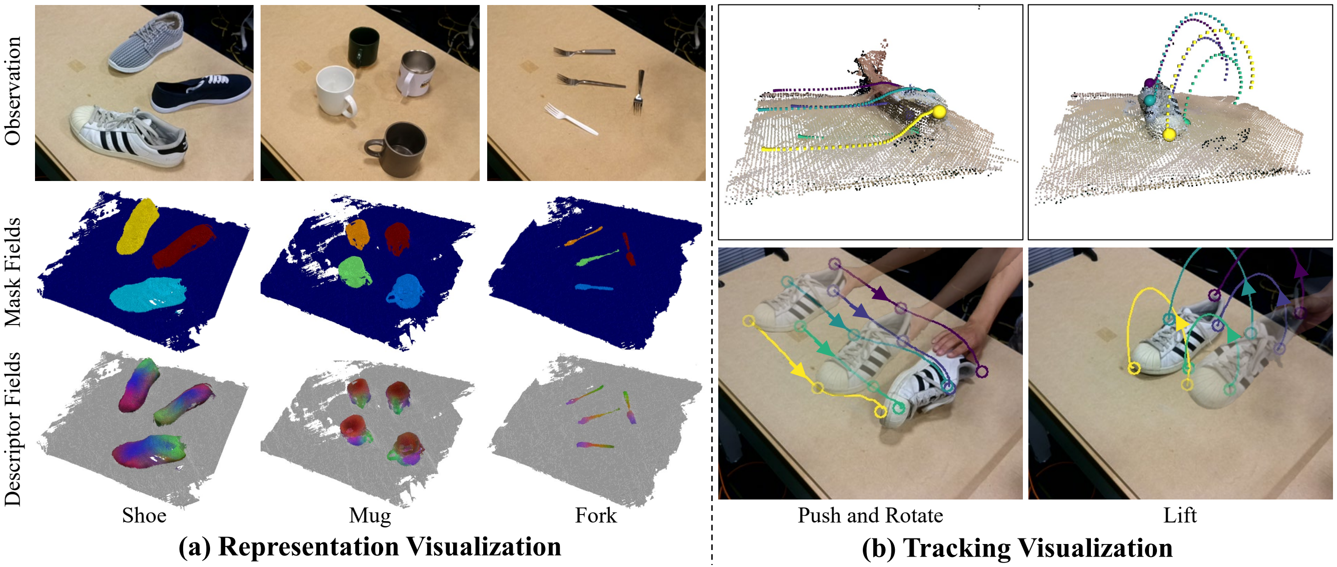
where Euclidean group $\mathbb{SE}(3) := \{\mathbf{R}, \mathbf{t} \mid \mathbf{R} \in \mathbb{SO}^3, \mathbf{t} \in \mathbb{R}^3\}$. For a 3D point $\mathbf{x}$ in the world frame, we could obtain projected pixel $\mathbf{u}_i$ and distance to camera $\mathbf{r}_i$ as follows:

$$\mathbf{u}_i = \pi\left(\mathbf{K}_i\left(\mathbf{R}_i\mathbf{x} + \mathbf{t}_i\right)\right), \quad \mathbf{r}_i = [0,0,1]^T\left(\mathbf{R}_i\mathbf{x} + \mathbf{t}_i\right), \tag{3}$$

where $\pi$ performs perspective projection, mapping a 3D vector $p = [x,y,z]^T$ to a 2D vector $q = [x/z, y/z]^T$.

### C. D$^3$Fields Representation

We fuse observation $\mathbf{o}^t$ from multiple views to build the implicit 3D descriptor fields $\mathcal{F}^t(\cdot)$. For simplicity, we will



**(a) Representation Visualization**

**(b) Tracking Visualization**

Fig. 3: **Representation and Tracking Visualizations.** (a) To verify that the representation is both 3D and semantic, we visualize the representation across different object categories. Mask fields color 3D points based on their instance masks, which clearly differentiates between instances. Descriptor fields color 3D points by mapping features to RGB space using PCA. They display a consistent color pattern within a category, such as mug handles being colorized as green for different mug instances. (b) To demonstrate that our representation is dynamic, we apply it to tracking tasks and showcase two tracking examples, both of which involve 3D motions and partial observations in single views. The robust 3D tracking results confirm that our representation is 3D, dynamic, and semantic.

represent $\mathbf{o}^t$ as $\mathbf{o}$, and $\mathcal{F}^t(\cdot)$ as $\mathcal{F}(\cdot)$ in this subsection. The implicit 3D descriptor field $\mathcal{F}(\cdot)$ is defined as

$$(\mathbf{d}, \mathbf{f}, \mathbf{p}) = \mathcal{F}(\mathbf{x}), \tag{4}$$

where $\mathbf{x}$ is an arbitrary 3D point in the world frame, and $(\mathbf{d}, \mathbf{f}, \mathbf{p})$ is the corresponding geometric and semantic descriptor. $\mathbf{d} \in \mathbb{R}$ is the signed distance from $\mathbf{x}$ to the surface. $\mathbf{f} \in \mathbb{R}^N$ represents the semantic information of $N$ dimension. $\mathbf{p} \in \mathbb{R}^M$ denotes the instance probability distribution of $M$ instances. $M$ could be different across scenarios.

More specifically, we denote a single view RGBD observation from camera $i$ as $\boldsymbol{o}_i = (\mathcal{I}_i, \mathcal{R}_i)$, where RGB image $\mathcal{I}_i \in \mathbb{R}^{H\times W\times 3}$, and depth image $\mathcal{R}_i \in \mathbb{R}^{H\times W}$. For an arbitrary 3D point $\mathbf{x}$, we project it to image space using Eq. 3 and use bilinear interpolation to obtain the corresponding depth $\mathbf{r}'_i = \mathcal{R}_i[\mathbf{u}_i]$. Then the descriptors from camera $i$ are

$$\begin{aligned} \mathbf{d}_i &= \max(\min(\mathbf{r}'_i - \mathbf{r}, \mu), -\mu), \\ \mathbf{f}_i &= \mathcal{W}_i^{\mathbf{f}}[\mathbf{u}_i], \quad \mathbf{p}_i = \mathcal{W}_i^{\mathbf{p}}[\mathbf{u}_i], \end{aligned} \tag{5}$$

where DINOv2 [17] extracts the semantic feature volume $\mathcal{W}_i^{\mathbf{f}} \in \mathbb{R}^{H\times W\times N}$ from RGB observation $\mathcal{I}_i$. $\mathcal{W}_i^{\mathbf{p}} \in \mathbb{R}^{H\times W\times M}$ is the instance mask volume using Grounded-SAM [14, 15]. $\mu$ is the truncation threshold for TSDF.

We fuse descriptors from all $K$ views as follows:

$$v_i = H(\mathbf{d}_i + \mu), \quad w_i = \exp\left(\frac{\min\left(\mu - |\mathbf{d_i}|, 0\right)}{\mu}\right), \tag{6}$$

and then

$$\mathbf{d} = \frac{\sum_{i=1}^K v_i \mathbf{d}_i}{\delta + \sum_{i=1}^K v_i}, \mathbf{f} = \frac{\sum_{i=1}^K v_i w_i \mathbf{f}_i}{\delta + \sum_{i=1}^K v_i}, \mathbf{m} = \frac{\sum_{i=1}^K v_i w_i \mathbf{m}_i}{\delta + \sum_{i=1}^K v_i}, \tag{7}$$

where $H$ is the unit step function and $\delta$ is a small value to avoid numeric errors. $v_i = 0$ when $\mathbf{x}$ is not observable in camera $i$, because if $\mathbf{x}$ is occluded in camera $i$, it should not contribute to the descriptor of $\mathbf{x}$. In addition, we could only have a confident estimation when $\mathbf{x}$ is close to the surface. Therefore, $w_i$ will decay as $|\mathbf{d}_i|$ increases. For $\mathbf{x}$ that is far away, $\mathbf{f}$ and $\mathbf{m}$ will degrade to $0^T$.

We convert the implicit field function $\mathcal{F}(\cdot)$ to a set of keypoints $\boldsymbol{s}$. First, we create voxels $\mathbf{x} \in \mathbb{R}^{W\times L\times H\times 3}$ in the workspace and evaluate $(\mathbf{d}, \mathbf{f}, \mathbf{p}) = \mathcal{F}(\mathbf{x})$. We filter out $\mathbf{x}_i \in \mathbf{x}$ where $\mathbf{d}_i$ is large or $\mathbf{p}_i$ has a low probability to avoid empty space and the background. After obtaining filtered points $\mathbf{x}'$, we use farthest point sampling to find surface points $\boldsymbol{s} \in \mathbb{R}^{3\times n_s}$ of an instance.

### D. Keypoints Tracking and Dynamics Training

This section will present how to use the dynamic implicit 3D descriptor field $\mathcal{F}(\cdot)$ to track keypoints and train dynamics. Without losing generalization, consider the tracking of a single instance $\boldsymbol{s}^t \in \mathbb{R}^{3\times n_s}$. For clarity, we denote $\mathbf{f}$ and $\mathbf{d}$ from $\mathcal{F}(\cdot)$ as $\mathcal{F}_{\mathbf{f}}(\cdot)$ and $\mathcal{F}_{\mathbf{d}}(\cdot)$. We formulate the tracking problem as an optimization problem:

$$\min_{\boldsymbol{s}^{t+1}} \quad ||\mathcal{F}_{\mathbf{f}}(\boldsymbol{s}^{t+1}) - \mathcal{F}_{\mathbf{f}}(\boldsymbol{s}_0)||_2. \tag{8}$$

Since $\mathcal{F}(\cdot)$ is differentiable, we could use a gradient-based optimizer. This method could be naturally extended to multiple-instance scenarios. We found that relying solely on features for tracking is unstable. We added rigid constraints and distance regularization for a more stable tracking.

Keypoint tracking enables dynamics model training on real data. We instantiate the dynamics model $f(\cdot, \cdot)$ as graph neural networks (GNNs). We follow [59] to predict object dynamics. Please refer to [25, 59] for more details on how to train the GNN-based dynamics model. The trained dynamics will be used for trajectory optimization in Section III-E.

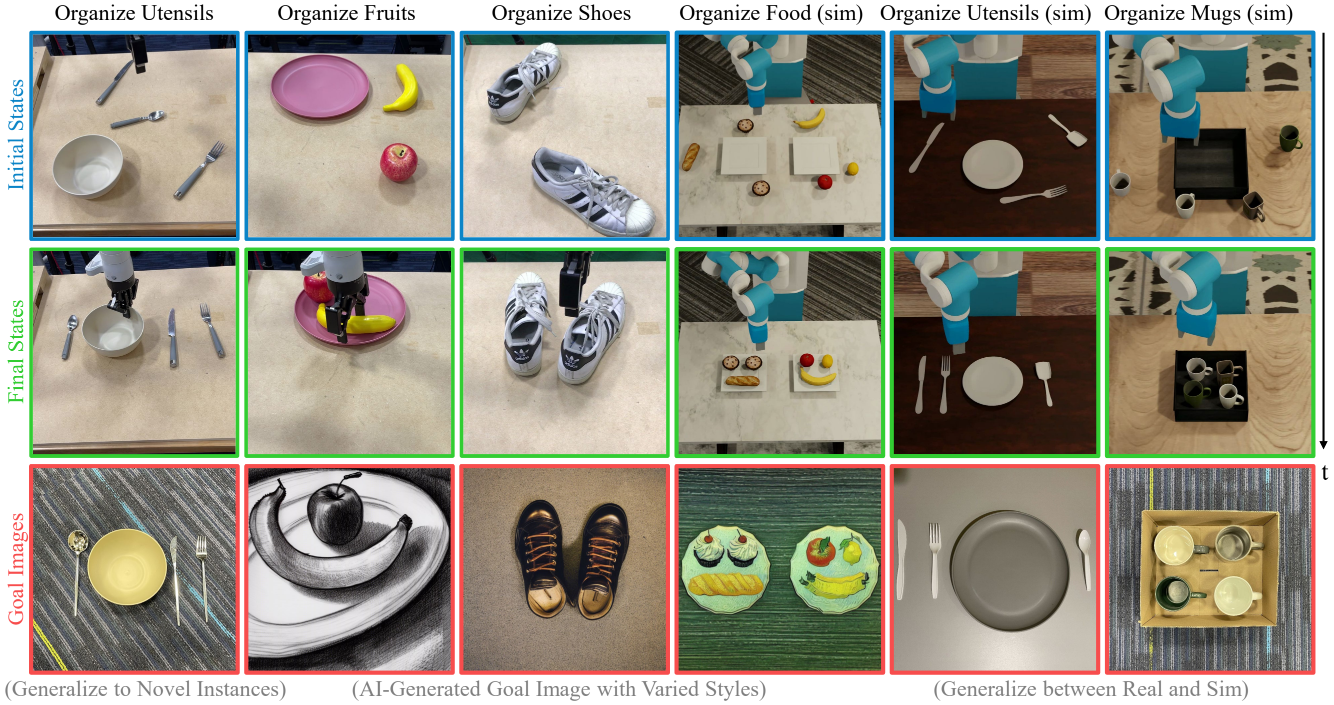


Fig. 4: **Qualitative results.** We qualitatively evaluate our proposed framework on household manipulation tasks, both in the real world and in simulation, encompassing tasks such as organizing utensils, fruits, shoes, food, and mugs. The figure highlights that our representation can generalize across varied instances, styles, and contexts. For instance, in the organizing fruits example, the goal image, unlike the workspace, is styled as a sketch drawing. Because our representation can map bananas with varied styles and appearances to similar features, the banana in the workspace can correspond to the banana in the sketch. This allows the task to be successfully completed. This wide range of tasks showcases the generalization capabilities and manipulation precision of our framework.

### *E. Zero-Shot Generalizable Robotic Manipulation*

As described in Section III-C, we denote initial tracked points and features as $\boldsymbol{s}^0$ and $\mathbf{f}^0$. We estimate $\boldsymbol{s}_{\text{goal}} \in \mathbb{R}^{2\times n_s}$ of goal image $\mathcal{I}_{\text{goal}}$ as follows:

$$\begin{aligned} \alpha_{ij} &= \exp\left(||\mathcal{W}^{\mathbf{f}}_{\text{goal}}[\mathbf{u}_i] - \mathbf{f}^0_j||_2\right), \\ w_{ij} &= \frac{\exp\left(s\alpha_{ij}\right)}{\sum_{i=1}^{H\times W}\exp\left(s\alpha_{ij}\right)}, \end{aligned} \tag{9}$$

then we have $\boldsymbol{s}_{\text{goal},j} = \sum_{i=1}^{H\times W} w_{ij}\mathbf{u}_i$, where $\mathcal{W}^{\mathbf{f}}_{\text{goal}}$ is the feature volume extracted from $\mathcal{I}_{\text{goal}}$ using DINOv2. $s$ is the hyperparameter to determine whether the heatmap $w_{ij}$ is more smooth or concentrating. Although Eq. 9 only shows a single instance case, it could be naturally extended to multiple instances by using instance mask information.

However, $\boldsymbol{s}_{\text{goal}}$ is in the image space, while $\boldsymbol{s}^t$ is in the 3D space. We bridge this gap by introducing a reference camera with approximate intrinsic and extrinsic parameters $\mathbf{K}'$ and $\mathbf{T}'$. Instead of rendering images in the reference view, we focus on projecting 3D keypoints into 2D images and define the task cost function in image space as follows:

$$c(\boldsymbol{s}^t, \boldsymbol{s}_{\text{goal}}) = ||\pi\left(\mathbf{K}'\left(\mathbf{R}'\boldsymbol{s}^t + \mathbf{t}'\right)\right) - \boldsymbol{s}_{\text{goal}}||_2^2. \tag{10}$$

## IV. Experiments

In this section, we evaluate our representation across various manipulation tasks with varying goal image styles, instances, and contexts. We visualize D$^3$Fields and showcase tracking results in Section IV-B. Then, we highlight our framework's zero-shot generalizability in both real-world and simulated tasks in Section IV-C. Finally, a quantitative comparison with baselines in Section IV-D underscores our framework's generalization and manipulation precision.

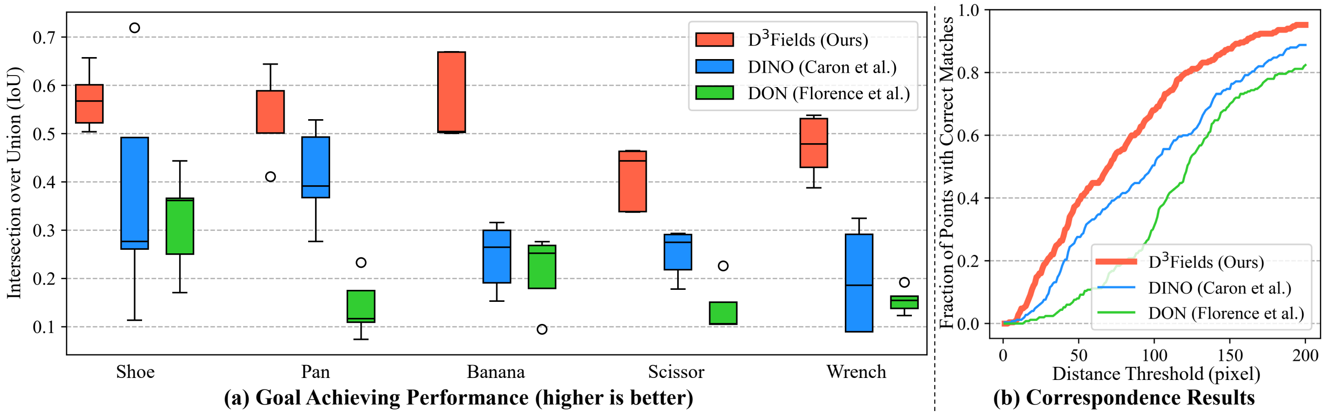
### *A. Experiment Setup*

In the real world, we employ four OAK-PRO D cameras to gather RGBD observations and use the Kinova® Gen3 for action execution. In simulation, we utilize OmniGibson and deploy Fetch for mobile manipulation tasks [60]. Our evaluations span a variety of tasks, including organizing shoes, collecting debris, tidying the office table, arranging utensils, and more.

We implement the baseline methods using Dense Object Nets (DON) and DINO for feature extraction [54, 61]. We quantitatively evaluate these methods on five object classes for single-instance manipulation tasks in the real world. The results and analysis are presented in Section IV-D.

### *B. Descriptor Fields Visualization and Keypoints Tracking*

D$^3$Fields provide a good 3D semantic representation, as shown in Fig. 3(a). We first visualize the mask fields by coloring 3D points according to their most likely instance, and our visualization shows a clear 3D instance segmentation. Additionally, we map the semantic features to RGB space using PCA, as with DINOv2 [17]. Visualization of the descriptor fields reveals that D$^3$Fields retain a dense semantic understanding of objects. In the provided shoe example, even



**(a) Goal Achieving Performance (higher is better)**

**(b) Correspondence Results**

Fig. 5: **Quantitative Evaluation.** We perform real-world quantitative evaluations by measuring final goal-achieving performance and keypoints correspondence accuracy. (a) We use IoU to measure goal-achieving performance. Results indicate that our method aligns with the goal configurations much better than DON and DINO across various object categories and scenarios. (b) We measure the keypoints correspondence accuracy according to the fraction of points with accurate matches, with correct matches determined by a distance threshold. Our method is consistently better at aligning with the goal image, regardless of the chosen threshold.

though various shoes have distinct appearances and poses, they exhibit similar color patterns: shoe heels are represented in green, and shoe toes in red. We observed similar patterns when evaluating the model on mugs and forks.

As discussed before, D$^3$Fields can also capture scene dynamics. We evaluate it by tracking the object keypoints. We show two examples of 3D keypoint tracking in Fig. 3(b). In the first example, a shoe is pushed and then flipped. Although only a portion of the shoe is visible from the view, our framework tracks it reliably. In another example, a shoe is lifted and then set down. Despite parts of the shoe being out of the camera's view, we can robustly track it in 3D.

### C. *Zero-Shot Generalizable Manipulation*

We conduct a qualitative evaluation of D$^3$Fields in common household robotic manipulation tasks in a zero-shot manner, with partial results displayed in Fig. 1 and Fig. 4. The following capabilities of our framework are observed:

**Generalization to AI-Generated Goal Images.** In Fig. 1, the goal image, rendered in a Van Gogh style, depicts shoes distinct from those in the workspace. Since D$^3$Fields encode semantic information, capturing shoes with varied appearances under similar descriptors, our framework can manipulate shoes based on AI-generated goal images.

**Compositional Goal Images and 3D Manipulation.** Using the office desk organization example in Fig. 1, the robot first arranges the mouse and pen according to the goal image. It then repositions the mug from the box to the mug pad, referencing a goal image of the upright mug.

**Generalization across Instances and Materials.** Granular objects, unlike rigid ones, have more complex dynamics. Our framework effectively handles these materials, as shown in the debris collection in Fig. 1. Fig. 4 further showcases our framework's instance-level generalization, where the goal image displays instances different from the workspace.

**Generalization across Simulation and Real World.** We evaluated our framework on household tasks in the simulator, as shown in the utensil organization and mug organization examples in Fig. 4. Given goal images taken from the real world, our framework can also manipulate objects to the goal configurations. Our framework demonstrates generalization capabilities between simulation and the real world.

### D. *Quantitative Comparisons with Baselines*

In Fig. 5(a), we measure performance using the IoU between the goal image mask and the final state mask after manipulation, with higher values indicating better alignment. Evaluating across five object classes, our method consistently outperforms the baselines, underscoring its generalization and manipulation accuracy. While DINO struggles with distinguishing object components, leading to imprecise results, it still works better than DON. Although DON performs well on familiar object classes and configurations, it lacks generalization in novel scenarios.

In Fig. 5(b), we present the correspondence results. We manually label corresponding keypoints on both the goal image and the final manipulation result to evaluate the correspondence accuracy. We calculate the fraction of accurately matched points based on a distance threshold. Our method consistently outperforms the baselines, regardless of the threshold. DINO ranks second, while DON lags behind. Consistent with Fig. 5(a), our method excels in generalization and accuracy, DINO is broadly applicable but less precise, and DON struggles with generalization.

## V. Conclusion

In this work, we introduce D$^3$Fields, which implicitly encode 3D semantic features and 3D instance masks, and model the underlying dynamics. Our emphasis is on zero-shot generalizable robotic manipulation tasks specified by 2D goal images of varying styles, contexts, and instances. Our framework excels in executing a diverse array of household manipulation tasks in both simulated and real-world scenarios. Its performance greatly surpasses baseline methods such as Dense Object Nets and DINO in terms of generalization capabilities and manipulation accuracy.

## References

[1] L. Manuelli, Y. Li, P. Florence, and R. Tedrake, "Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning," in *Conference on Robot Learning (CoRL)*, 2020.

[2] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba, "3d neural scene representations for visuomotor control," *arXiv preprint arXiv:2107.04004*, 2021.

[3] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "Robocraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks," *arXiv preprint arXiv:2205.02909*, 2022.

[4] Y. Ze, N. Hansen, Y. Chen, M. Jain, and X. Wang, "Visual reinforcement learning with self-supervised 3d representations," *IEEE Robotics and Automation Letters (RA-L)*, 2023.

[5] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.

[6] W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, "Learning predictive representations for deformable objects using contrastive estimation," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 564–574.

[7] Y. Wang, Y. Li, K. Driggs-Campbell, L. Fei-Fei, and J. Wu, "Dynamic-Resolution Model Learning for Object Pile Manipulation," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[8] M. Minderer, C. Sun, R. Villegas, F. Cole, K. P. Murphy, and H. Lee, "Unsupervised learning of object structure and dynamics from videos," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning*. PMLR, 2018, pp. 306–316.

[10] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 13 081–13 088.

[11] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, G. Iyer, S. Saryazdi, T. Chen, A. Maalouf, S. Li, N. V. Keetha, A. Tewari, J. Tenenbaum, C. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, "ConceptFusion: Open-set multimodal 3D mapping," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[12] K. Mazur, E. Sucar, and A. J. Davison, "Feature-realistic neural fusion for real-time, open set scene understanding," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8201–8207.

[13] W. Liu, Y. Du, T. Hermans, S. Chernova, and C. Paxton, "Structdiffusion: Language-guided creation of physically-valid structures using unseen objects," in *RSS 2023*, 2023.

[14] 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 preprint arXiv:2303.05499*, 2023.

[15] 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," *arXiv:2304.02643*, 2023.

[16] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *ECCV*, 2022.

[17] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[18] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[19] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *Conference on Robot Learning*. PMLR, 2023, pp. 1769–1782.

[20] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[21] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *7th Annual Conference on Robot Learning*, 2023.

[22] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *7th Annual Conference on Robot Learning*, 2023.

[23] K. Mülling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 263–279, 2013.

[24] Y. Duan, M. Andrychowicz, B. Stadie, O. Jonathan Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," *Advances in neural in-*

*formation processing systems*, vol. 30, 2017.

[25] Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, and A. Torralba, "Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids," in *ICLR*, 2019.

[26] W. Wang, A. S. Morgan, A. M. Dollar, and G. D. Hager, "Dynamical scene representation and control with keypoint-conditioned neural radiance field," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 1138–1143.

[27] 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.

[28] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kpam: Keypoint affordances for category-level robotic manipulation," in *The International Symposium of Robotics Research*. Springer, 2019, pp. 132–157.

[29] W. Gao and R. Tedrake, "kpam-sc: Generalizable manipulation planning using keypoint affordance and shape completion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6527–6533.

[30] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[31] X. Lin, C. Qi, Y. Zhang, Y. Li, Z. Huang, K. Fragkiadaki, C. Gan, and D. Held, "Planning with spatial-temporal abstraction from point clouds for deformable object manipulation," in *6th Annual Conference on Robot Learning*, 2022.

[32] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 892–909.

[33] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv:2203.06173*, 2022.

[34] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," *CoRL*, 2022.

[35] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran, and V. Kumar, "Cacti: A framework for scalable multi-task multi-scene visual imitation learning," *arXiv preprint arXiv:2212.05711*, 2022.

[36] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, and K. Hausman, "Open-world object manipulation using pre-trained vision-language models," 2023.

[37] Y. Yoon, G. N. DeSouza, and A. C. Kak, "Real-time tracking and pose estimation for industrial objects using geometric features," in *2003 IEEE International conference on robotics and automation (cat. no. 03CH37422)*, vol. 3. IEEE, 2003, pp. 3473–3478.

[38] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M. Zhang, C. Phillips, M. Lecce, and K. Daniilidis, "Single image 3d object detection and pose estimation for grasping," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 3936–3943.

[39] L. Zhu, A. Mousavian, Y. Xiang, H. Mazhar, J. van Eenbergen, S. Debnath, and D. Fox, "Rgb-d local implicit function for depth completion of transparent objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4649–4658.

[40] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-nerf: Using a neural radiance field to grasp transparent objects," in *5th Annual Conference on Robot Learning*, 2021.

[41] Y. Wi, P. Florence, A. Zeng, and N. Fazeli, "Virdo: Visio-tactile implicit representations of deformable objects," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3583–3590.

[42] 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 *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6394–6400.

[43] D. Driess, J.-S. Ha, M. Toussaint, and R. Tedrake, "Learning models as functionals of signed-distance fields for manipulation planning," in *Conference on Robot Learning*. PMLR, 2022, pp. 245–255.

[44] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations," in *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[45] T. Weng, D. Held, F. Meier, and M. Mukadam, "Neural grasp distance fields for robot manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[46] D. Driess, I. Schubert, P. Florence, Y. Li, and M. Toussaint, "Reinforcement learning with neural radiance fields," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[47] D. Shim, S. Lee, and H. J. Kim, "SNeRL: Semantic-aware neural radiance fields for reinforcement learning," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 31 489–31 503.

[48] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola, "NeRF-Supervision: Learning dense object descriptors from neural radiance fields," in *IEEE Conference on Robotics and Automation (ICRA)*, 2022.

[49] Z. Tang, B. Sundaralingam, J. Tremblay, B. Wen, Y. Yuan, S. Tyree, C. Loop, A. Schwing, and S. Birchfield, "RGB-only reconstruction of tabletop scenes for collision-free manipulator control," in *ICRA*, 2023.

[50] A. Zhou, M. J. Kim, L. Wang, P. Florence, and C. Finn, "Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 907–17 917.

[51] N. M. (Mahi)Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "CLIP-Fields: Weakly Supervised Semantic Fields for Robotic Memory," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[52] Y. Wi, A. Zeng, P. Florence, and N. Fazeli, "Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects," *arXiv preprint arXiv:2210.03701*, 2022.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[54] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[55] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *International Conference on Computer Vision (ICCV)*, 2023.

[56] S. Sharma, A. Rashid, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *7th Annual Conference on Robot Learning*, 2023.

[57] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot manipulation," in *7th Annual Conference on Robot Learning*, 2023.

[58] Y. Ze, G. Yan, Y.-H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L. E. Li, and X. Wang, "Multi-task real robot learning with generalizable neural feature fields," in *7th Annual Conference on Robot Learning*, 2023.

[59] Y. Li, T. Lin, K. Yi, D. Bear, D. L. Yamins, J. Wu, J. B. Tenenbaum, and A. Torralba, "Visual grounding of learned physical models," in *International Conference on Machine Learning*, 2020.

[60] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, M. Anvari, M. Hwang, M. Sharma, A. Aydin, D. Bansal, S. Hunter, K.-Y. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, S. Savarese, H. Gweon, K. Liu, J. Wu, and L. Fei-Fei, "BEHAVIOR-1k: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation," in *6th Annual Conference on Robot Learning*, 2022.

[61] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," in *Proceedings of The 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, 29–31 Oct 2018, pp. 373–385.