Affordances from Human Videos as a Versatile Representation for Robotics

Shikhar Bahl[⋆1,2] Russell Mendonca[⋆1] Lili Chen[1] Unnat Jain[1,2] Deepak Pathak[1]

[1]CMU [2]Meta AI

Figure 1. We leverage human videos to learn visual affordances that can be deployed on multiple real robot, in the wild, spanning several tasks and learning paradigms. Videos available at https://vision-robotics-bridge.github.io/.

Abstract

Building a robot that can understand and learn to interact by watching humans has inspired several vision problems. However, despite some successful results on static datasets, it remains unclear how current models can be used on a robot directly. In this paper, we aim to bridge this gap by leveraging videos of human interactions in an environment centric manner. Utilizing internet videos of human behavior, we train a visual affordance model that estimates where *and* how *in the scene a human is likely to interact. The structure of these behavioral affordances directly enables the robot to perform many complex tasks. We show how to seamlessly integrate our affordance model with four robot learning paradigms including offline imitation learning, exploration, goal-conditioned learning, and action parameterization for reinforcement learning. We show the efficacy of our approach, which we call Vision-Robotics Bridge (VRB) across 4 real world environments, over 10 different tasks, and 2 robotic platforms operating in the wild.*

> *The meaning or value of a thing consists of what it affords... what we perceive when we look at objects are their affordances, not their qualities.*
>
> *J.J. Gibson (1979)*

1. Introduction

Imagine standing in a brand-new kitchen. Before taking even a single action, we already have a good understanding of how most objects should be manipulated. This understanding goes beyond semantics as we have a belief of where to hold objects and which direction to move them in, allowing us to interact with it. For instance, the oven is opened by pulling the handle downwards, the tap should be turned sideways, drawers are to be pulled outwards, and light switches are turned on with a flick. While things don't always work as imagined and some exploration might be needed, but humans heavily rely on such visual *affordances* of objects to efficiently perform day-to-day tasks across environments [34, 35]. Extracting such actionable knowledge from videos has long inspired the vision community.

More recently, with improving performance on static datasets, the field is increasingly adopting a broader 'active' definition of vision through research in egocentric visual understanding and visual affordances from videos of human interaction. With deep learning, methods can now predict heatmaps of where a human would interact [38, 75] or seg-

[⋆]equal contribution

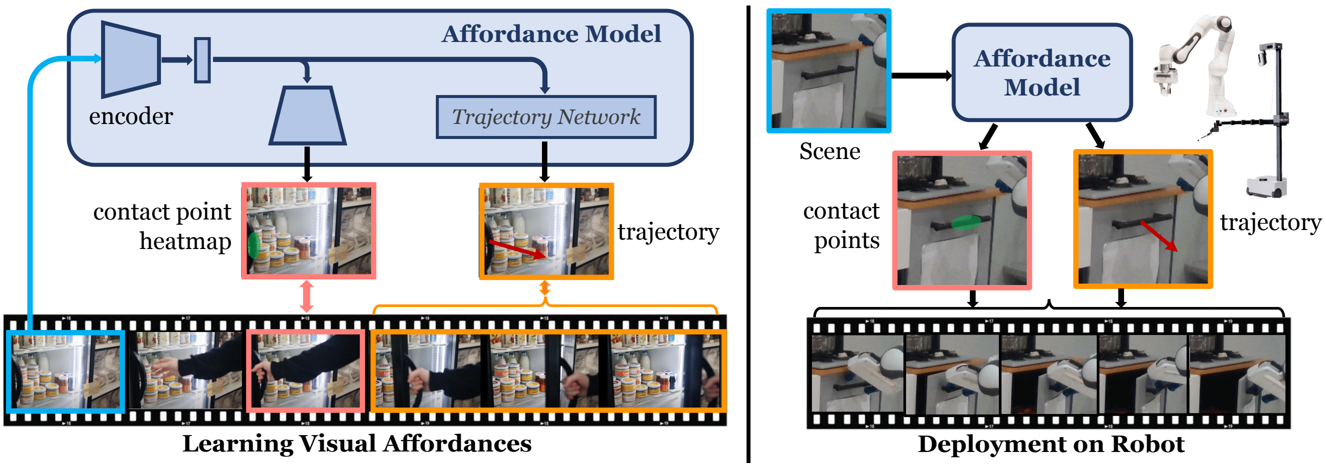


Figure 2. **VRB Overview**. First, we learn an actionable representation of visual affordances from human videos: the model predicts contact points and trajectory waypoints with supervision from future frames. For robot deployment, we query the affordance model and convert its outputs to 3D actions to execute.

mentation of the object being interacted with [101]. Despite being motivated by the goal of enabling downstream robotic tasks, prior methods for affordance learning are tested primarily on human video datasets with no physical robot or in-the-wild experiments. Without integration with a robotic system, even the most basic question of how the affordance should be defined or represented remains unanswered, let alone evaluating its performance.

On the contrary, most robot learning approaches, whether imitation or reinforcement learning, approach a new task or a new environment *tabula rasa*. At best, the visual representation might be pretrained on some dataset [65, 79, 91, 100, 115, 117]. However, visual representations are only a small part of the larger problem. In robotics, especially in continuous control, the state space complexity grows exponentially with actions. Thus, even with perfect perception, knowing what to do is difficult. Given an image, current computer vision approaches can label most of the objects, and even tell us approximately where they are but this is not sufficient for the robot to perform the task. It also needs to know *where* and how to manipulate the object, and figuring this out from scratch in every new environment is virtually impossible for all but the simplest of tasks. How do we alleviate this clear gap between visual learning and robotics?

In this paper, we propose to rethink visual affordances as a means to bridge vision and robotics. We argue that rich video datasets of humans interacting can offer a lot more actionable information beyond just replacing ImageNet as a pretrained visual encoder for robot learning. Particularly, human interactions are a rich source of how a wide range of objects can be held and what are useful ways to manipulate their state. However, several challenges hinder the smooth integration of vision and robotics. We group them into three parts. *First*, what is an actionable way to represent affordances? *Second*, how to learn this representation in a data-driven and scalable manner? *Third*, how to adapt visual affordances for deployment across robot learning paradigms? To answer the first question, we find that contact points and post-contact trajectories are excellent robot-centric representations of visual affordances, as well as modeling the inherent multi-modality of possible interactions. We make effective use of egocentric datasets in order to tackle the second question. In particular, we reformulate the data to focus on frames without humans for predicting contact points and the post-contact trajectories. To extract free supervision for this prediction, we utilize off-the-shelf tools for estimating egomotion, human pose, and hand-object interaction. Finally, we show how to seamlessly integrate these affordance priors with different kinds of robot learning paradigms. We call our approach **V**ision-**R**obotics **B**ridge (VRB) due to its core goal of bridging vision and robotics.

We evaluate both the quality of our affordances and their usefulness for 4 different robotic paradigms – imitation and offline learning, exploration, visual goal-reaching, and using the affordance model as a parameterization for action spaces. These are studied via extensive and rigorous real-world experiments on physical robots which span across 10 real-world tasks, 4 environments, and 2 robot hardware platforms. Many of these tasks are performed *in-the-wild* outside of lab environments (see Figure 1). We find that VRB outperforms other state-of-the-art human hand-object affordance models, and enables high-performance robot learning in the wild without requiring any simulation. Finally, we also observe that our affordance model learns a good visual representation for robotics as a byproduct. We highlight that all the evaluations are **performed in the real world spanning several hundred hours of robot running time** which is a very large-scale evaluation in robotics.

2. Related Work

Affordance and Interaction Learning from Videos. Given a scene, one can predict interactions using geometry-based rules for objects via 3D scene understanding [42, 74, 127], estimating 3D physical attributes [8, 25, 40, 130] or through segmentation models trained on semantic interactions [97, 98]. These approaches, however, require specialized datasets. More general interaction information can be learned from large human datasets [18–20, 39, 59, 63], to predict object information [29, 129] (RGB & 3D) [10], graphs [23] or environment information [27, 76] such as heatmaps [38, 75]. Approaches also track human poses, especially hands [14, 18, 61, 62, 96, 101, 120]. Similarly, in action anticipation and human motion forecasting, high-level semantic or low level actions are predicted using visual history [1, 11, 19, 21, 30–32, 36, 39, 44–46, 52, 55, 68, 71, 95, 112, 113]. Since our observations only have robot arms and no human hands, we adopt a robot-first formulation, only modeling the contact point and post-contact phase of interaction.

Visual Robot Learning. Learning control from visual inputs directly is an important challenge. Previous works have leveraged spatial structures of convolutional networks to directly output locations for grasping and pushing from just an image of the scene [87, 123, 124], which can limit the type of tasks possible. It is also possible to directly learn control end-to-end [50, 58] which while general, is quite sample inefficient in the real world. It has been common to introduce some form of prior derived from human knowledge, which could take the form of corrective interactions [22, 41, 64], structured policy spaces [2, 7, 7, 17, 48, 80, 89, 94, 102, 118], offline robotics data [24, 53, 54, 67, 92], using pretrained visual representations [79, 84, 100, 116, 117] or human demonstrations [6, 15, 99, 102, 103, 107].

Learning Manipulation from Humans. Extensive work has been done on Learning from Demonstrations (LfD) where human supervision is usually provided through teleoperation (of a joystick or VR interface) [73, 109, 126] or kinesthetic teaching, where a user physically moves the robot arm [13, 16, 26, 66, 89]. With both these approaches, collecting demonstrations is tedious and slow. Recently, works have shown alternate ways to provide human demonstrations, via hand pose estimation and retargeting [5, 90, 104, 106, 119] in robot hands, but are mostly restricted to tabletop setups. First and third person human demonstrations have been used to train policies directly, transferred either via a handheld gripper [82, 108, 121] or using online adaptation [6]. In contrast to directly mimicking a demonstration, we learn robot-centric *affordances* from passive human videos that provide a great initialization for downstream robot tasks, unlike previous work which require in-domain demonstrations.

3. Vision-Robotics Bridge (VRB)

Our goal is to learn affordance priors from large-scale egocentric videos of human interaction, and then use them to expedite robot learning in the wild. This requires addressing the three questions discussed in Sec. 1 about how to best represent affordances, how to extract them and how to use them across robot learning paradigms.

3.1. Actionable Representation for Affordances

Affordances are only meaningful if there is an actor to execute them. For example, a chair has a sitting affordance only if it is possible for some person to sit on it. This property makes it clear that the most natural way to extract human affordances is by watching how people interact with the world. However, what is the right object-centric representation for affordances: is it a heatmap of where the human makes contact? Is it the pre and postcondition of the object? Is it a description of the human interaction? All of these are correct answers and have been studied in prior works [42, 62, 75]. However, the affordance parameterization should be amenable to deployment on robots.

If we want the robot to *a priori* understand how to manipulate a pan (Fig. 1, 4) without any interaction, then a seemingly simple solution is to exactly model human movement from videos [62], but this leads to a human-centric model and will not generalize well because human morphology is starkly different from that of robots. Instead, we take a first-principles approach driven by the needs of robot learning. Knowledge of a robot body is often known, hence reaching a point in the 3D space is feasible using motion planning [51, 56, 57]. The difficulty is in figuring out where to interact (e.g. the handle of the lid) and then how to move after the contact is made (e.g., move the lid upwards).

Inspired by this, we adopt contact points and post-contact trajectories as a simple actionable representation of visual affordance that can be easily transferred to robots. We use the notation c for a contact point and τ for post-contact trajectory, both in the pixel space. Specifically, $\tau = f(I_t, h_t)$, where I_t is the image at timestep t, h_t is the human hand location in pixel space, and f is a learned model. We find that our affordance representation outperforms prior formulations across robots. Notably, the c and τ abstraction makes the affordance prior agnostic to the morphological differences across robots.

3.2. Learning Affordances from Egocentric Videos

The next question is how to extract c and τ from human videos in a scalable data-driven manner while dealing with the presence of human body or hand in the visual input. VRB tackles this through a robot-first approach.

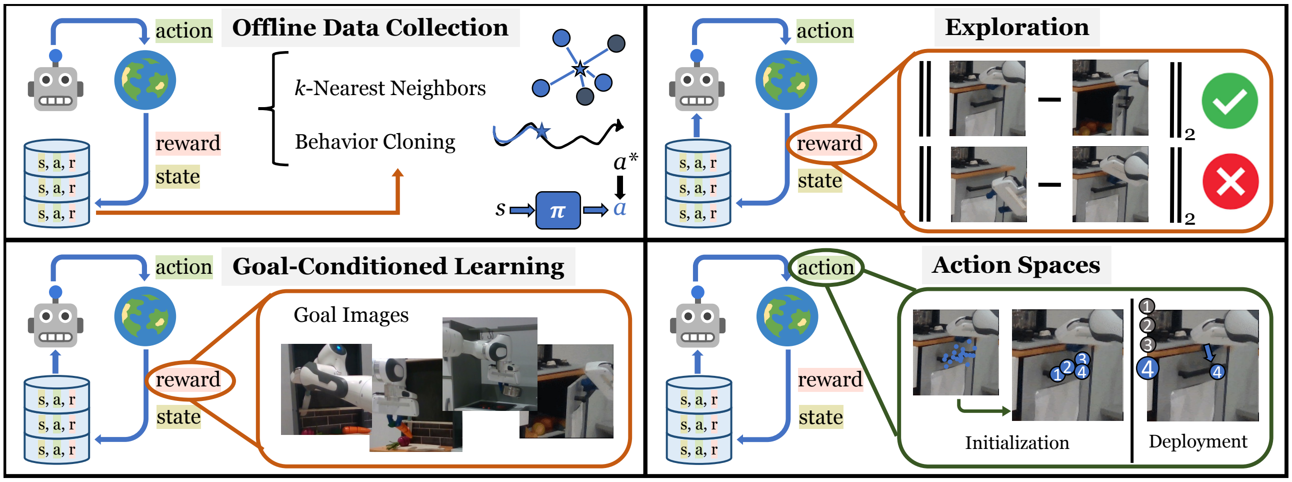


Figure 3. **Robot Learning Paradigms** : (a) Offline Data Collection – Used to investigate the quality of the collected data. (b) Exploration – The robot needs to use intrinsic rewards to improve (c) Goal-Conditioned Learning – A desired task is specified via a goal image, used to provide reward. (d) Action Spaces – Reduced action spaces are easier to search and allow for discrete control.

3.2.1 Extracting Affordances from Human Videos

Consider a video V, say of a person opening a door, consisting of T frames *i.e.* $V = \{I_1, ..., I_T\}$. We have a twofold objective — find *where* and *when* the contact happened, and estimate how the hand moved after contact was made. This is used to supervise the predictive model $f_\theta(I_t)$ that outputs contact points and post-contact trajectories. To do so, we utilize a widely-adopted hand-object detection model trained on human video data [101]. For each image I_t, this produces 2D bounding boxes of the hand h_t, and a discrete contact variable o_t. Using this information, we filter for frames where o_t indicates a contact in each video, and find the first timestep where contact occurs, t_{contact}.

The pixel-space positions of the hand $\{h_t\}_{t_{\text{contact}}}^{t'}$ constitute the post-contact trajectory (τ). To extract contact points c, we use the corresponding hand bounding box, and apply skin color segmentation to find all points at the periphery of the hand segment that intersect with the bounding box of the object in contact. This gives us a set of N contact points $\{c^i\}^N$, where N can differ depending on the image, object, scene and type of interaction. How should the contact points be aggregated to train our affordance model (f_θ)? Some options include predicting the mean of $\{c^i\}^N$, or randomly sampling c^i. However, we seek to encourage multi-modality in the predictions, since a scene likely contains multiple possible interactions. To enable this, we fit a Gaussian mixture model (GMM) to the points. Let us define a distribution over contact points to be $p(c)$. We fit the GMM parameters (μ_k, Σ_k) and weights α_k.

$$p(c) = \underset{\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k \mathcal{N}(c^i|\mu_k, \Sigma_k) \quad (1)$$

We use these parameters of the above defined GMM with K clusters as targets for f_θ. To summarize, 1) we find the first timestep where contact occurs in the human video, t_{contact} 2) For c, we fit a GMM to the contact points around the hand at frame $I_{t_{\text{contact}}}$, parameterized by μ_k, Σ_k and 3) we find the post-contact trajectory of the 2D hand bounding box $\{h_t\}_{t_{\text{contact}}}^{t'}$ for τ.

Accounting for Camera Motion over Time: Consider a person opening a door. Not only do the person's hands move but their body and hence their head also move closer to the handle and then away from it. Therefore, we need to compensate for this egomotion of the human head/camera from time t_{contact} to t'. We address this by using the homography matrix at timestep t, $\mathcal{H}_t$ to project the points back into the coordinates of the starting frame. We obtain the homography matrix by matching features between consecutive frames. We then use this to produce the transformed trajectory $\tau = \mathcal{H}_t \circ \{h_t\}_{t_{\text{contact}}}^{t'}$.

Addressing Human-Robot Visual Domain Shift: The training videos contain human body or hand in the frame but the human will not be present in downstream robotics task, generating domain shift. We deal with this issue with a simple yet elegant trick: we extract affordances in the frames with humans but then map those affordances back to the first frame when human was yet to enter the scene. For videos in which a human is always in frame, we either crop out the human in the initial frame if there is no interaction yet or discard the frame if the human is always in contact. We compute the contact points and post-contact trajectories with respect to this human-less frame via the same homography procedure described above. This human-less frame is then used to condition our affordance model.

3.2.2 Training Affordance Model

Conditioned on the input image, the affordance model is trained to predict the extracted labels for contact points and post-contact trajectories. However, naive joint prediction does not work well as the learning problem is inherently multi-modal. For instance, one would pick up a cup differently from a table depending on whether the goal is to pour it into the sink or take a sip from it. We handle this by predicting multiple heatmaps for interaction points using the same model, building a spatial probability distribution.

For ease of notation, we use $(\cdot)_\theta$ as a catch-all for all parameterized modules and use f_θ to denote our complete network. Fig. 2 shows an overview of our model. Input image I_t is encoded using a ResNet [43] visual encoder g_θ^{conv} to give a spatial latent representation z_t, i.e., $g_\theta^{\text{conv}}(I_t) = z_t$. We then project this latent z_t into K probability distributions or heatmaps using deconvolutional layers; concretely, $H_t = g_\theta^{\text{deconv}}(z_t)$. Using a spatial softmax, σ_{2D}, we get the estimation of the labels for GMM means, *i.e.*, μ_k. We found that keeping the covariance matrices fixed gave better results. Formally, the loss for contact point estimation is:

$$\mathcal{L}_{\text{contact}} = \left\| \mu_i - \sigma_{\text{2D}}\left(g_\theta^{\text{deconv}}\left(g_\theta^{\text{conv}}\left(I_t\right)\right)\right)\right\|_2 \qquad (2)$$

To estimate post-contact trajectory, we train a trajectory prediction network, $\mathcal{T}_\theta$, based on the latent representation z_t. We find that it is easier to optimize for *relative* shifts, *i.e.*, the direction of movement instead of absolute locations, assuming that the first point $\hat{w}_0$ is 0, since the contact points are already spatially grounded. Based on the success of Transformers for sequential prediction, we employ self-attention blocks [111] and train to optimize $\mathcal{L}_{\text{traj}} = \|\tau - \mathcal{T}_\theta(z_t)\|_2$. In a given scene, there are many objects a human could interact with, which may or may not be present in the training data. We tackle this uncertainty and avoid spurious correlations by sampling local crops of I_t around the contact points. These serve as the effective input to our network f_θ and enables better generalization.

3.3. Robot Learning from Visual Affordances

Instead of finding a particular way to use our affordance model for robotics, we show that it can bootstrap existing robot learning methods. In particular, we consider four different robotics paradigms as shown in Fig. 3.

A. Imitation Learning from Offline Data Collection Imitation learning is conventionally performed on data collected by human demonstrations, teleoperation, or scripted policies – all of which are expensive and only allow for small-scale data collection [4, 6, 12, 58, 103, 122]. On the other hand, using the affordance model, $f_\theta(\cdot)$ to guide the robot has a high probability of yielding 'interesting' interactions.

Given an image input I_t, the affordance model produces $(c, \tau) = f_\theta(I_t)$, and we store $\{(I_t, (c, \tau))\}$ in a dataset $\mathcal{D}$. After sufficient data has been collected, we can use imitation learning to learn control policies, often to complete a specific task. A common approach for task specification is to use *goal images* that show the desired configuration of objects. Given the goal image, the *k-Nearest Neighbors* (k-NN) approach involves filtering trajectories in $\mathcal{D}$ based on their distance to the goal image in feature space. Further, the top (filtered) trajectories can be used for *behavior cloning* (BC) by training a policy, $\pi(c, \tau | I_t)$, We run both k-NN and behavior cloning on datasets collected by different methods in Sec. 4.1. Using the same IL approach for different datasets is also useful for comparing the relative quality of the data. This is because higher relative success for a particular dataset implies that the data is qualitatively better, given that the same IL algorithm achieves worse performance on a different dataset. This indicates that the goal (or similar images) were likely seen during data collection.

B. Reward-Free Exploration The goal of exploration is to discover as many diverse skills as possible which can aid the robot in solving downstream tasks. Exploration methods are usually guided by *intrinsic rewards* that are self-generated by the robotic agent, and are not specific to any task [9, 47, 49, 60, 69, 81, 85, 88, 93, 110]. However, starting exploration from scratch is too inefficient in the real world, as the robot can spend an extremely large amount of time trying to explore and still not learn meaningful skills to solve tasks desired by humans. Here our affordance model can be greatly beneficial by bootstrapping the exploration from the predicted affordances allowing the agent to focus on parts of the scene likely to be of interest to humans. To operationalize this, we first use the affordance model $f_\theta(.)$ for data-collection. We then rank all the trajectories collected using a task-agnostic exploration metric, and fit a distribution h to the (c, τ) values of the top trajectories. For subsequent data collection, we sample from h with some probability, and otherwise use the affordance model f. This process can then be repeated, and the elite-fitting scheme will bootstrap from highly exploratory trajectories to improve exploration even further. For the exploration metric in our experiments, we maximize *environment change* $\text{EC}(I_i, I_j) = ||\phi(I_i) - \phi(I_j)||_2$, (similar to previous exploration approaches [6, 83]) between first and last images in the trajectory, where ϕ masks the robot and the loss is only taken on non-masked pixels.

C. Goal-Conditioned Learning While exploring the environment can lead to interesting skills, consider a robot that already knows its goal. Using this knowledge (*e.g.* an image of the opened door), it supervise its policy search. Goal images are frequently used to specify rewards in RL [3, 33, 37, 70, 77, 78, 86, 114, 131]. Using our affordance

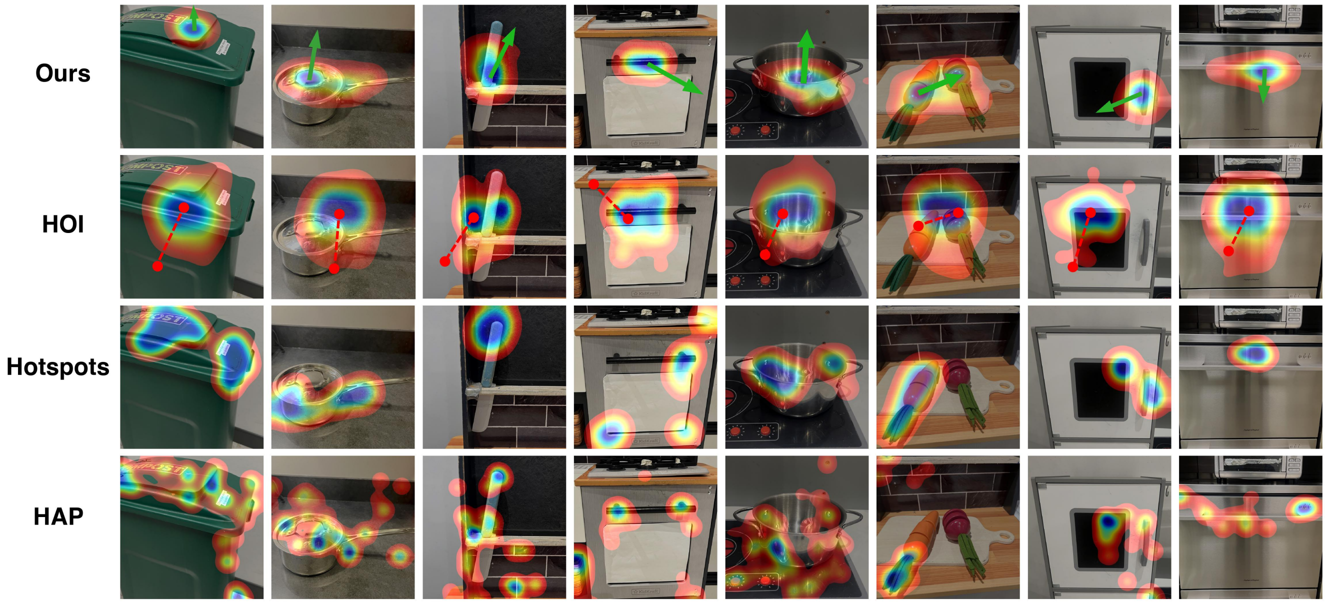


Figure 4. Qualitative affordance model outputs for `VRB`, `HOI` [62], `Hotspots` [38] and `HAP` [38], showing the predicted contact point region, and post-grasp trajectory (green arrow for `VRB`, red for `HOI` [62]). We can see that `VRB` produces the most meaningful affordances.

model can expedite the process of solving goal-specified tasks. Similar to the exploration setting, we rank trajectories and fit a distribution h to the (c, τ) values of the top trajectories, but here the metric is to minimize distance to the goal image I_g. The metric used in our experiments is to minimize $\mathrm{EC}(I_T, I_g)$, where I_T is the last image in the trajectory, or to minimize $||\psi(I_g) - \psi(I_T)||_2^2$, where ψ is a feature space. Akin to exploration, subsequent data collection involves sampling from h and the affordance model f.

D. Affordance as an Action Space Unlike games with discrete spaces like Chess and Go where reinforcement learning is deployed *tabula rasa*, robots need to operate in continuous action spaces that are difficult to optimize over. A pragmatic alternative to continuous action spaces is parameterizing them in a spatial manner and assigning a primitive (e.g. grasping, pushing or placing) to each location [105, 124, 125]. While this generally limits the type of tasks that can be performed, our affordance model already seeks out interesting states, due to the data it is trained on. We first query the affordance model on the scene many times to obtain a large number of predictions. We then fit a GMM to these points to obtain a discrete set of (c, τ) values, and now the robot just needs to search over this space.

4. Experimental Setup and Results

Through the four robot learning paradigms, shown in Fig. 3, we seek to answer the following questions: (1) Does our model enable a robot to collect *useful data* (imitation from offline data)?, (2) How much benefit does VRB provide to *exploration* methods?, (3) Can our method enable *goal-conditioned* learning?, and (4) Can our model be used to define a structured *action space* for robots? Finally, we also study whether our model learns meaningful *visual representations* for control as a byproduct and also analyze the *failure modes* and how they differ from prior work.

Robotics Setup We use two different robot platforms - the Franka Emika Panda arm and the Hello Stretch mobile manipulator. We run the Franka on two distinct play kitchen environments and test on tasks that involve interacting with a cabinet, a knife and some vegetables, and manipulation of a a shelf and a pot. The Hello robot is tested on multiple in-the wild tasks outside lab settings, including opening a garbage can, lifting a lid, opening a door, pulling out a drawer, and opening a dishwasher (Fig. 1). We also provide support for a simulation environment on the Franka-Kitchen benchmark [28]. Details can be found in the Appendix.

Observation and Action space For each task, we estimate a task-space image-crop using bounding boxes [128], and pass random sub-crops to f_θ. The prediction for contact points c and post-contact trajectory τ is in pixel space, which are projected into 3D for robot control using a calibrated robot-camera system (with an Intel RealSense D415i). The robot operates in 6DOF end-effector space – samples a rotation, moves to a contact point, grasps, and then moves to a post-contact position (see Sec. 3.1).

Baselines and Ablations: We compare against prior work that has tried to predict heatmaps from human video : 1) Hotspots [75] 2) Hands as Probes (`HAP`) [38], a modified version for our robot setup of Liu *et al*. [62] that predicts

	Cabinet	Knife	Veg	Shelf	Pot	Door	Lid	Drawer
k-Nearest Neighbors:								
`HOI`	0.2	0.1	0.1	0.6	0.0	0.4	0.0	0.6
`HAP`	0.3	0.0	0.3	0.0	0.1	0.2	0.0	0.1
`Hotspots`	0.4	0.0	0.1	0.0	**0.5**	0.4	0.3	0.5
`Random`	0.3	0.0	0.1	0.3	0.4	0.2	0.1	0.2
`VRB` (ours)	**0.6**	**0.3**	**0.6**	**0.8**	0.4	**1.0**	**0.4**	**1.0**
Behavior Cloning:								
`HOI`	0.3	0.0	0.3	0.0	0.1	0.2	0.0	0.1
`HAP`	0.5	0.0	**0.4**	0.0	0.3	0.1	0.0	0.1
`Hotspots`	0.2	0.0	0.0	0.0	**0.8**	0.1	0.0	0.7
`Random`	0.1	**0.1**	0.1	0.0	0.2	0.1	0.0	0.0
`VRB` (ours)	**0.6**	**0.1**	0.3	**0.3**	**0.8**	**0.9**	**0.2**	**0.9**

Table 1. **Imitation Learning**: Success rate for k-NN and Behavior Cloning on collected offline data using various affordance models. We find that VRB vastly outperforms prior approaches, indicating better quality of data.

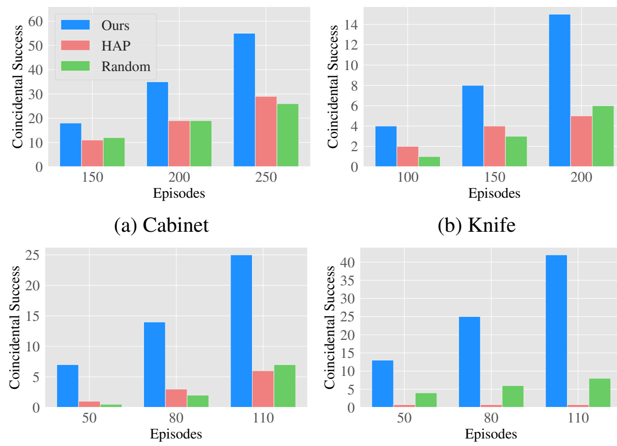
contact region and forecast hand poses: 3) `HOI` [62] and 4) a baseline that samples affordances at random (`Random`). HAP and Hotspots only output a contact point, and we randomly select a post-contact direction. More details are available in the Appendix.

4.1. Quality of Collected Data for Imitation

We investigate VRB as a tool for useful data collection. We evaluate this on both our robots across 8 different environments, with results in Tab. 1. These are all unseen scenarios (not in train set). Tasks are specified for each environment using goal images (eg - open door, lifted pot etc), and we use the data collected (30-150 episodes) for two established offline learning methods: (1) k-Nearest Neighbors (k-NN) and (2) Behavior Cloning. k-NN [82] finds trajectories in the dataset that are close (via distance in feature space [79]) to the goal image. We run the 10-closest trajectories to the goal image and record whether the robot has achieved the task specified in the goal image. For behavior cloning, we train a network supervised with (image, waypoint) pairs from the collected dataset, and the resulting policy is run 10 times on the real system. With both k-NN and BC, our method outperforms prior tasks on 7 out of 8 tasks, with an average success rate of 57 %, with the runner-up method (Hotspots [75]) only getting 25 %. This shows that VRB leads to much better data offline data quality, and thus can lead to better imitation learning performance. We additionally test for grasping held-out *rare* objects such as VR remotes or staplers, and find that VRB outperforms baselines. Details can be found in the Appendix.

4.2. Reward-Free Exploration

Here we study self-supervised exploration with no external rewards. We utilize environment change, *i.e.*, change in the position of objects as a task-agnostic metric for exploration [6]. For improved exploration, we bias sampling towards trajectories with a higher environment change metric. To evaluate the quality of exploration data, we measure how often does the robot achieves coincidental success *i.e.* reach a goal image configuration without having access to it. As shown in Fig. 5, we obtain consistent improvements over HAP [38] and random exploration raising performance multiple fold – from $3\times$ to $10\times$, for every task.



(a) Cabinet (b) Knife

(c) Stovepot (d) Shelf

Figure 5. **Exploration**: Coincidental success of VRB in comparison to random exploration or the exploration based on HAP [38].

4.3. Goal-Conditioned Learning

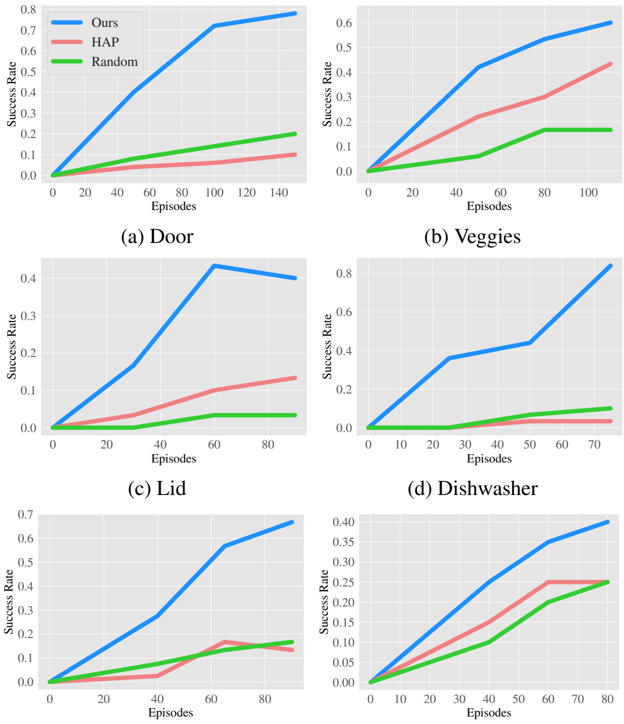
The previous settings help robots improve their behaviors with data without an external reward or goal. Here we focus on goal-driven robot learning. Goals are often specified through images of the goal configuration. Note that goal images are also used in Sec. 4.1 but as part of a static dataset to imitate. Here, the robot policy is updated with new data being added to the buffer. We sample this dataset for trajectories that minimize visual change with respect to the goal image. As shown in Fig. 6, VRB learns faster and better HAP [38] and Random on this robot learning paradigm, over six diverse tasks.

4.4. Affordance as an Action Space

We utilize visual affordances to create a discrete action space using a set of contact points and post-contact trajectories. We then train a Deep Q-Network (DQN) [72] over this action space, for the above goal-conditioned learning problem. In Fig. 7, we see that with VRB, the robot experiences more successes showing that a greater percentage of actions in the discretized action space correspond to meaningful object interactions.

4.5. Analyzing Visual Representations

Beyond showing better utility for robot learning paradigms, we analyze the quality of visual representations of the encoder learned in VRB. Two standard evaluations

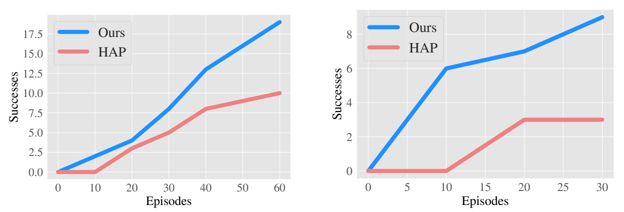


(a) Door (b) Veggies

(c) Lid (d) Dishwasher

(e) Drawer (f) Garbage Can

Figure 6. **Goal-conditioned Learning**: Success rate for reaching goal configuration for six different tasks. Sampling via VRB leads to faster learning and better final performance.



(a) Cabinet (b) Veggies

Figure 7. **Action Space**: Success using DQN with the discretized action space, for reaching a specified goal image.

for this are (1) if they can help for downstream tasks and (2) how meaningful distances in their feature spaces are.

	VRB	R3M
microwave	**0.16**	0.10
slide-door	**0.84**	0.70
door-open	**0.13**	0.11

Table 2. Behavior Cloning with VRB *vs*. R3M [79] representation.

Finetuning To investigate if the visual representations are effective for control, we directly finetune a policy on top of the (frozen) visual encoder. We evaluate on three simulated Franka environments, as shown in Tab. 2, and we see that VRB outperforms R3M on all tasks. (We finetuned the policy only for 2K steps, instead of 20K in the R3M paper). This demonstrates that VRB visual representations contain information that is useful for control.

Feature space distance We record the distance in feature space between the current and goal image for every timestep

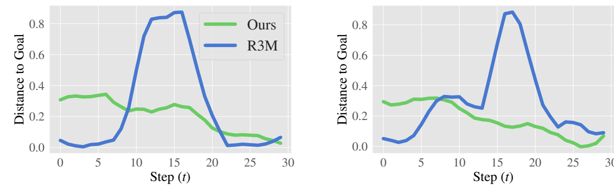


Figure 8. **Feature space distance**: Distance to goal in feature space for VRB decreases monotonically for door opening.

in the episode, for both VRB and R3M [79] on successful cabinet opening trajectories. As shown in Fig. 8, the distance for VRB decreases almost monotonically which correlates well with actual task progress.

4.6. Failure Modes

While VRB and the baselines see qualitatively similar successes, VRB in general sees a larger number of them and the *average case* scenario for VRB is much better. For the cabinet opening task, we classify each collected episode into three categories: "Failure", "Partial Success" and "Success". While VRB has a higher number of successful trajectories compared to the baselines (almost $2\times$), the number of partial successes is more than $6\times$ (Fig. 9).

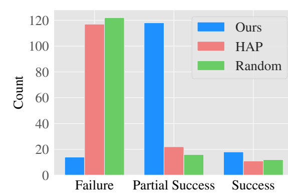


Figure 9. Failure mode analysis

5. Conclusion

We propose Vision-Robotics Bridge (VRB), a scalable approach for learning useful affordances from passive human video data, and deploying them on many different robot learning paradigms (such as data collection for imitation, reward-free exploration, goal conditioned learning and paramterizing action spaces). Our affordance representation consists of contact points and post-contact trajectories. We demonstrate the effectiveness of this approach on the four paradigms and 10 different real world robotics tasks, including many that are in the wild. We run thorough experiments, spanning over 200 hours, and show that VRB drastically outperforms prior approaches. In the future, we hope to deploy on more complex multi-stage tasks, incorporate physical concepts such as force and tactile information, and investigate VRB in the context of visual representations.

Acknowledgements We thank Shivam Duggal, Yufei Ye and Homanga Bharadhwaj for very fruitful discussions and are grateful to Shagun Uppal, Ananye Agarwal, Murtaza Dalal and Jason Zhang for comments on early drafts of this paper. RM, LC and DP are supported by ONR MURI N00014-22-1-2773 and ONR N00014-22-1-2096.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *CVPR*, 2018. 3

[2] Brandon Amos, Ivan Dario Jimenez Rodriguez, Jacob Sacks, Byron Boots, and J. Zico Kolter. Differentiable mpc for end-to-end planning and control. In *NeurIPS*, 2018. 3

[3] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *NIPS*, 2017. 5

[4] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 2009. 5

[5] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. *arXiv preprint arXiv:2203.13251*, 2022. 3

[6] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *RSS*, 2022. 3, 5, 7

[7] Shikhar Bahl, Mustafa Mukadam, Abhinav Gupta, and Deepak Pathak. Neural dynamic policies for end-to-end sensorimotor learning. In *NeurIPS*, 2020. 3

[8] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *CVPR*, 2016. 3

[9] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *NIPS*, 2016. 5

[10] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. First person action-object detection with egonet. *arXiv preprint arXiv:1603.04908*, 2016. 3

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 3

[12] Arthur E Bryson and Yu-Chi Ho. *Applied optimal control: optimization, estimation, and control*. Routledge, 2018. 5

[13] Sylvain Calinon, Florent Guenter, and Aude Billard. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 2007. 3

[14] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 3

[15] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021. 3

[16] Vivian Chu, Tesca Fitzgerald, and Andrea L Thomaz. Learning object affordances by leveraging the combination of human-guidance and self-exploration. In *International Conference on Human-Robot Interaction*, 2016. 3

[17] Murtaza Dalal, Deepak Pathak, and Russ R Salakhutdinov. Accelerating robotic reinforcement learning via parameterized action primitives. *NeurIPS*, 2021. 3

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022. 3

[19] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 3

[20] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *NeurIPS Track on Datasets and Benchmarks*, 2022. 3

[21] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017. 3

[22] Todor Davchev, Kevin Sebastian Luck, Michael Burke, Franziska Meier, Stefan Schaal, and Subramanian Ramamoorthy. Residual learning from demonstration. *arXiv preprint arXiv:2008.07682*, 2020. 3

[23] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *TPAMI*, 2021. 3

[24] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021. 3

[25] D Eigen and R Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. corr, abs/1411.4734. *arXiv preprint arXiv:1411.4734*, 2014. 3

[26] Sarah Elliott, Zhe Xu, and Maya Cakmak. Learning generalizable surface cleaning actions from demonstration. In *International Symposium on Robot and Human Interactive Communication*, 2017. 3

[27] David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV*, 2012. 3

[28] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020. 6

[29] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 2017. 3

[30] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *ICCV*, 2019. 3

[31] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *TPAMI*, 2020. 3

[32] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 3

[33] Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. *arXiv preprint arXiv:1912.06088*, 2019. 5

[34] JJ Gibson. The ecological approach to visual perception. *Houghton Mifflin Comp*, 1979. 1

[35] James Jerome Gibson. *The senses considered as perceptual systems*, volume 2. 1

[36] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021. 3

[37] Ankit Goyal, Arsalan Mousavian, Chris Paxton, Yu-Wei Chao, Brian Okorn, Jia Deng, and Dieter Fox. Ifor: Iterative flow minimization for robotic object rearrangement. In *CVPR*, 2022. 5

[38] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *CVPR*, 2022. 1, 3, 6, 7

[39] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 3

[40] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 3

[41] Reymundo A Gutierrez, Vivian Chu, Andrea L Thomaz, and Scott Niekum. Incremental task modification via corrective demonstrations. In *ICRA*, 2018. 3

[42] M Hassanin, S Khan, and M Tahtali. Visual affordance and function understanding: a survey. arxiv. *arXiv preprint arXiv:1807.06775*, 2018. 3

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[44] De-An Huang and Kris M Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *ECCV*, 2014. 3

[45] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *ICRA*, 2016. 3

[46] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. In *CVPR*, 2018. 3

[47] Unnat Jain, Iou-Jen Liu, Svetlana Lazebnik, Aniruddha Kembhavi, Luca Weihs, and Alexander G Schwing. Gridtopix: Training embodied agents with minimal supervision. In *ICCV*, 2021. 5

[48] Unnat Jain, Luca Weihs, Eric Kolve, Ali Farhadi, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander G. Schwing. A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In *ECCV*, 2020. 3

[49] Unnat Jain, Luca Weihs, Eric Kolve, Mohammad Rastegari, Svetlana Lazebnik, Ali Farhadi, Alexander G. Schwing, and Aniruddha Kembhavi. Two body problem: Collaborative visual task completion. In *CVPR*, 2019. 5

[50] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018. 3

[51] Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *IJRR*, 2011. 3

[52] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *TPAMI*, 2015. 3

[53] Aviral Kumar, Anikait Singh, Stephen Tian, Chelsea Finn, and Sergey Levine. A workflow for offline model-free robotic reinforcement learning. *arXiv preprint arXiv:2109.10813*, 2021. 3

[54] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020. 3

[55] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *ECCV*, 2014. 3

[56] Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006. 3

[57] S. M. Lavalle and J. J. Kuffner. Rapidly-exploring random trees: Progress and prospects. *Algorithmic and Computational Robotics: New Directions*, 2000. 3

[58] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 2016. 3, 5

[59] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018. 3

[60] Iou-Jen Liu, Unnat Jain, Raymond A Yeh, and Alexander Schwing. Cooperative exploration for multi-agent deep reinforcement learning. In *ICML*, 2021. 5

[61] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *ECCV*, 2020. 3

[62] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, 2022. 3, 6, 7

[63] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022. 3

[64] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, 2022. 3

[65] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: To-

wards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022. 2

[66] Guilherme J Maeda, Gerhard Neumann, Marco Ewerton, Rudolf Lioutikov, Oliver Kroemer, and Jan Peters. Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks. *Autonomous Robots*, 2017. 3

[67] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021. 3

[68] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting@ ego4d challenge 2022. *arXiv preprint arXiv:2207.12080*, 2022. 3

[69] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Alan: Autonomously exploring robotic agents in the real world. In *ICRA*, 2023. 5

[70] Lina Mezghan, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *IROS*, 2022. 5

[71] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. In *NeurIPS*, 2022. 3

[72] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb. 2015. 7

[73] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L Sidner, and Daniel Miller. Interactive hierarchical task learning from a single demonstration. In *International Conference on Human-Robot Interaction*, 2015. 3

[74] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015. 3

[75] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 1, 3, 6, 7

[76] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 3

[77] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. *ICRA*, 2017. 5

[78] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *NeurIPS*, pages 9191–9200, 2018. 5

[79] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 2, 3, 7, 8

[80] Soroush Nasiriany, Huihan Liu, and Yuke Zhu. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In *ICRA*, 2022. 3

[81] Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. Count-based exploration with neural density models. *ICML*, 2018. 5

[82] Jyothish Pari, Nur Muhammad, Sridhar Pandian Arunachalam, Lerrel Pinto, et al. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021. 3, 7

[83] Simone Parisi, Victoria Dean, Deepak Pathak, and Abhinav Gupta. Interesting object, curious agent: Learning task-agnostic exploration. *Advances in Neural Information Processing Systems*, 34:20516–20530, 2021. 5

[84] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. *arXiv preprint arXiv:2203.03580*, 2022. 3

[85] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017. 5

[86] Deepak Pathak, Parsa Mahmoudieh, Guanghao Luo, Pulkit Agrawal, Dian Chen, Yide Shentu, Evan Shelhamer, Jitendra Malik, Alexei A Efros, and Trevor Darrell. Zero-shot visual imitation. In *ICLR*, 2018. 5

[87] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *ICRA*, 2016. 3

[88] Vitchyr H Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019. 5

[89] M. Prada, A. Remazeilles, A. Koene, and S. Endo. Dynamic movement primitives for human-robot interaction: Comparison with human behavioral observation. In *IROS*, 2013. 3

[90] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, 2022. 3

[91] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022. 2

[92] Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In *Learning for Dynamics and Control*, 2021. 3

[93] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *IJCV*, 2021. 5

[94] Nathan D Ratliff, Jan Issac, Daniel Kappler, Stan Birchfield, and Dieter Fox. Riemannian motion policies. *arXiv preprint arXiv:1801.02854*, 2018. 3

[95] Nicholas Rhinehart and Kris M Kitani. Learning action maps of large environments via first-person vision. In *CVPR*, pages 580–588, 2016. 3

[96] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021. 3

[97] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016. 3

[98] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *CVPR*, 2017. 3

[99] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *ICRA*, 2018. 3

[100] Rutav M Shah and Vikash Kumar. Rrl: Resnet as representation for reinforcement learning. In *ICML*, 2021. 2, 3

[101] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020. 2, 3, 4

[102] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *IJRR*, 2021. 3

[103] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. *arXiv preprint arXiv:1911.09676*, 2019. 3, 5

[104] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *CoRL*, 2022. 3

[105] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *CoRL*, 2022. 6

[106] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. *RSS*, 2022. 3

[107] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. In *RSS*, 2020. 3

[108] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *RAL*, 2020. 3

[109] Jordi Spranger, Roxana Buzatoiu, Athanasios Polydoros, Lazaros Nalpantidis, and Evangelos Boukas. Human-machine interface for remote training of robot tasks. *arXiv preprint arXiv:1809.09558*, 2018. 3

[110] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *NIPS*, 2017. 5

[111] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 5

[112] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 3

[113] Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. Predicting motivations of actions by leveraging text. In *CVPR*, 2016. 3

[114] Justin Wasserman, Karmesh Yadav, Girish Chowdhary, Abhinav Gupta, and Unnat Jain. Last-mile embodied visual navigation. In *CoRL*, 2022. 5

[115] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv:2203.06173*, 2022. 2

[116] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 3

[117] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. *arXiv:2204.13226*, 2022. 2, 3

[118] Mengyuan Yan, Gen Li, Yilin Zhu, and Jeannette Bohg. Learning topological motion primitives for knot planning. In *IROS*, 2020. 3

[119] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *arXiv preprint arXiv:2207.05053*, 2022. 3

[120] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 3

[121] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. *arXiv preprint arXiv:2008.04899*, 2020. 3

[122] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *RSS*, 2018. 5

[123] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. *CoRL*, 2020. 3

[124] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *IROS*, 2018. 3, 6

[125] Kuo-Hao Zeng, Luca Weihs, Ali Farhadi, and Roozbeh Mottaghi. Pushing it out of the way: Interactive visual navigation. In *CVPR*, 2021. 6

[126] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *ICRA*, 2018. 3

[127] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *CVPR*, 2013. 3

[128] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 6

[129] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 3

[130] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *CVPR*, 2016. 3

[131] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017. 5