Human Hands as Probes for Interactive Object Understanding

Mohit Goyal Sahil Modi Rishabh Goyal Saurabh Gupta
University of Illinois Urbana-Champaign
{mohit, smodi9, rgoyal6, saurabhg}@illinois.edu

Abstract

Interactive object understanding, or what we can do to objects and how is a long-standing goal of computer vision. In this paper, we tackle this problem through observation of human hands in in-the-wild egocentric videos. We demonstrate that observation of what human hands interact with and how can provide both the relevant data and the necessary supervision. Attending to hands, readily localizes and stabilizes active objects for learning and reveals places where interactions with objects occur. Analyzing the hands shows what we can do to objects and how. We apply these basic principles on the EPIC-KITCHENS *dataset, and successfully learn state-sensitive features, and object affordances (regions of interaction and afforded grasps), purely by observing hands in egocentric videos.*

1. Introduction

Consider the cupboard in Figure 1. Merely localizing and naming it is insufficient for a robot to successfully interact with it. To enable interaction we, we need to identify what are plausible sites for interaction, how should we interact with each site, and what would happens when we do. The goal of this paper is to acquire such an understanding about objects. Specifically, we formulate it as a) learning a feature space that is sensitive to the *state* of the object (and thus indicative of what we can do with it) rather than just its *category*; and b) identifying what hand-grasps do objects afford and where. These together provide an interactive understanding of objects, and could aid learning policies for robots. For instance, distance in a state-sensitive feature space can be used as reward functions for manipulation tasks [52,54,64]. Similarly, hand-grasps afforded by objects and their locations provide priors for exploration [38, 39].

While we have made large strides in building models for how objects look (the various object recognition problems), the same recipe of collecting large-scale labeled datasets for training doesn't quite apply for understanding how objects work. First of all, no large-scale labeled datasets already exist for such tasks. Second, manually annotating these aspects on static images is challenging. For instance, objects states are highly contextual: the same object (*e.g.* cupboard in Figure 1) can exist in many different states (closed, full, on-top-of, has-handle, in-contact-with-hand) at the same time, depending on the interaction we want to conduct. Similarly, consciously annotating where and how one can touch an object can suffer from biases, leading to data that may not be indicative of how people *actually* use objects during normal daily conduct. While one might annotate that we pull on the handle to open the cupboard; in real life we may very often just flick it open by sliding our fingers in between the cupboard door and its frame.

Motivated by these challenges, we pursue learning directly from the *natural* ways in which people interact with objects in egocentric videos. Since, egocentric data focuses upon hand-object interaction, it solves both the data and the supervision problem. Egocentric observation of human hands reveals information about the objects they are inter-

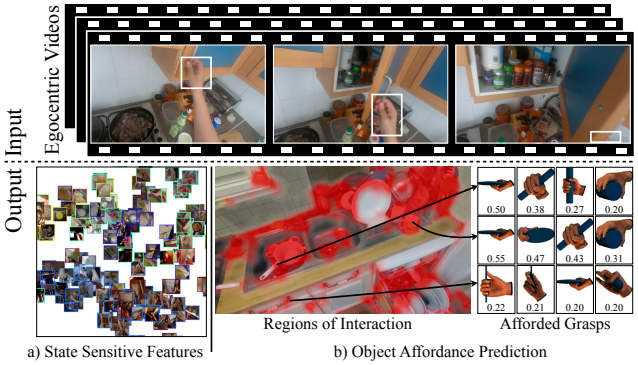

Figure 1. Human hands reveal information about objects as they interact with them. They tell us where and how we can interact with an object (the handle of the cupboard via an adducted thumb grasp), and what happens when we do (cupboard opens to reveals many more objects within). This paper develops techniques to extract an interactive understanding of objects through the observation of hands in a corpus of egocentric videos. Specifically, we produce a) features that are indicative of object states, and b) object affordances (*i.e.* regions of interaction, and afforded grasps).

Project website: https://s-gupta.github.io/hands-as-probes/.

acting with. Attending to locations that hands attend to, localizes and stabilizes active objects in the scene for learning. It shows where all hands can interact in the scene. Analyzing what the hand is doing reveals information about the state of the object, and also how to interact with it. Thus, observation of human hands in egocentric videos can provide the necessary data and supervision for obtaining an interactive understanding of objects.

To realize these intuitions, we design novel techniques that extract an understanding of objects from the understanding of hands as obtained from off-the-shelf models. We apply this approach to the two aspects of interactive object understanding: a) learning state-sensitive features, and b) inferring object affordances (identifying what hand-grasps do objects placed in scenes afford and where).

For the former goal of learning state-sensitive features, we *hand-stabilize* the object-of-interaction. We exploit the appearance and motion of the hand as it interacts with the object to derive supervision for the object state. This is done through contrastive learning where we encourage objects associated with similar hand appearance and motion, to be similar to one another. This leads to features that are more state-sensitive than those obtained from alternate forms of self-supervision, and even direct semantic supervision.

For the latter goal of predicting regions-of-interaction and applicable grasps, we additionally use hand grasp-type predictions. As the hand is directly visible when the interaction is happening, the challenge here is to get the model to focus on the object to make its predictions, rather than the hand. For this, we design a context prediction task: we mask-out the hand and train a model to predict the location and grasp-type from the surrounding context. We find that modern models can successfully learn to make such contextual predictions. This enables us to identify the places where humans interact in scenes. We better recall small interaction sites such as knobs and handles, and also make more specific predictions when interaction sites are localized to specific regions on the objects (*e.g.* knobs for stoves). We are also able to successfully learn hand-grasps applicable to different objects.

For both these aspects, deriving supervision from hands sidesteps the need for and possible pitfalls of semantic supervision. We are able to conduct learning without having to define a complete taxonomy of object states, or suffer from inherent ambiguity in defining action classes.

2. Related Work

We survey research on understanding human hands, using humans or their hands as cues, interactive object understanding, and self-supervision.

Understanding hands. Several works have sought to build a data-driven understanding of human hands and how they manipulate objects from RGB images [63], RGB-D images [51], egocentric data [31], videos [17] and other sensors [2, 58]. This understanding can take different forms: grasp type classification [3, 51, 63] from a hand-defined taxonomy [15], hand keypoint and pose estimation [17], understanding gestures [18], detecting hands, their states and objects of interaction [55, 56], 3D reconstruction of the hand and the object of interaction [4, 21], or even estimating forces being applied by the hand onto the object [13]. We refer the reader to the survey paper from Bandini and Zariffa [1] for an analysis of hand understanding in context of egocentric data. Our goals are different: we build upon the understanding of hands to better understand objects.

Using humans or their hands as probes. The most relevant research to our work is that of using humans and hands as probes for understanding objects, scenes and other humans. [16, 57, 61] learn about scene affordance by watching how people interact with scenes in videos from YouTube, sitcoms and self-driving cars. Brahmbhatt *et al*. [2] learn task-oriented grasping regions by analyzing where people touch objects using thermal imaging. Wang *et al*. [60] use humans as visual cues for detecting novel objects. Mandikal and Grauman [38] extend work from [2] to learn policies for object manipulation using predicted contact regions. Ng *et al*. [45] use body pose of another person to predict the self-pose in egocentric videos. Unlike these past works, we focus upon observation of hands (and not full humans) in unscripted in-the-wild RGB egocentric videos (rather than in lab or with specialized sensors), to learn fine-grained aspects of object affordance (rather than scene affordance). Concurrent work from Nagarajan *et al*. [43] works in a similar setting but focuses on learning activity-context priors.

Interactive Object Understanding. Observing hands interact with objects is not the only way to learn about how to interact with objects. Researchers have used other forms of supervision (strong supervision, weak supervision, imitation learning, reinforcement learning, inverse reinforcement learning) to build interactive understanding of objects. This can be in the form of learning a) where and how to grasp [9, 20, 25, 26, 28, 33, 34, 37, 41, 49], b) state classifiers [24], c) interaction hotspots [14, 40, 42, 59], d) spatial priors for action sites [44], e) object articulation modes [11, 36], f) reward functions [27, 29, 48, 50], g) functional correspondences [32]. While our work pursues similar goals, we differ in our supervision source (observation of human hands interacting with objects in egocentric videos).

Self-supervision. Our techniques are inspired by work in self-supervision where the goal is to learn without semantic labels [5, 7, 8, 12, 19, 65]. Specifically, our work builds upon recent use of context prediction [12, 47] and contrastive learning [6, 53] for self-supervision. We design novel sources of supervision in the context of egocentric videos to enable interactive object understanding.

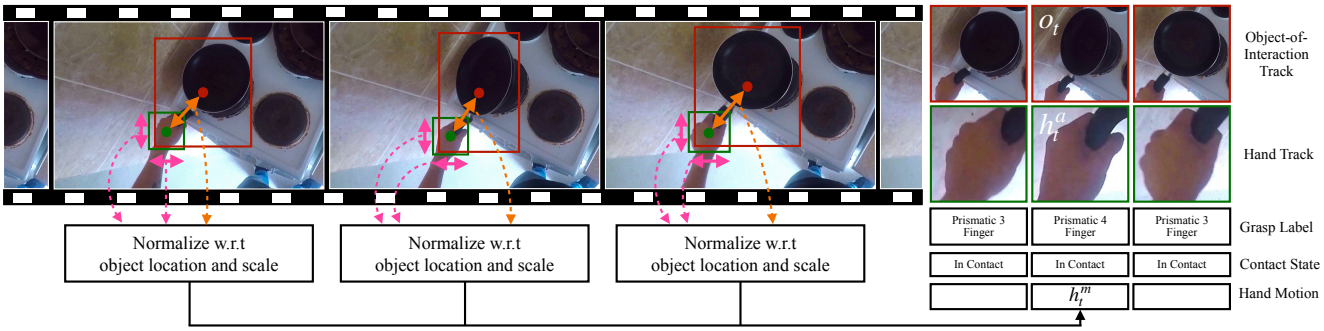

Figure 2. Data preparation. Given egocentric videos from the EPIC-KITCHENS dataset [10], we obtain per-frame detections for hand, object-of-interaction, and contact-state from [55]. These detections are strung together over time to form paired object-hand tracks. We represent the motion of the hand h_t^m around the object by stacking the hand box location and scale relative to the object over 3 adjacent frames. **Object-of-interaction tracks, hand tracks and hand motion** are together used to learn state sensitive feature spaces (Section 3.1). We also obtain hand grasp labels through a classifier trained on the GUN-71 dataset [51]. **The detected hand and object-of-interaction pairs along with these hand grasp labels** are used for learning regions of interaction and the grasps afforded by these regions (Section 3.2).

3. Approach

We work with the challenging EPIC-KITCHENS dataset from Damen *et al*. [10], and use the hand and object-of-interaction detector from Shan *et al*. [55] This detector provides per-frame detection boxes for both hands and the objects undergoing interaction, along with the hand contact state (whether the hands are touching something or not). We further obtain predictions for hand grasp-types for the detected hands, using a model trained on the 71-way grasp-type classification dataset from Rogez *et al*. [51]. We string together detected hands and objects-of-interaction in consecutive frames to form object-of-interaction and hand tracks as shown in Figure 2. We use these tracks for learning state-sensitive features (Section 3.1). Affordances (where and how hands interact with objects) are learned using per-frame predictions (Section 3.2).

3.1. State Sensitive Features via Temporal and Hand Consistency

Our formulation builds upon two key ideas: consistency of object states *in time* and *with hand pose*. Our training objective encourages object crops, that are close in time or are associated with similar hand appearance and motion, to be similar to one another; while being far from random other object crops in the dataset. We realize this intuition through contrastive learning and propose a joint loss: $L_{\text{temporal}} + \lambda L_{\text{hand}}$. L_{temporal} encourages *temporal* consistency by sampling naturally occurring temporal augmentations as additional transforms. L_{hand} uses hands as contrasting examples; positives being the hands that temporally correspond to the object crop, and negatives being other randomly sampled hands. L_{hand} indirectly encourages similarity between *different* objects that are similarly interacted by hands, and so are likely to be in similar states.

We construct batches for contrastive learning by sampling an object crop o_i and a temporally close hand crop h_i^a from tracks shown in Figure 2. We also encode the hand motion h_i^m, by concatenating the location and scale of the hand box relative to the object box over three neighboring frames. h_i^a and h_i^m jointly represent the hand: h_i^a describes the appearance and h_i^m describes the motion. We sample another frame o_i' from the same object track, as a temporal augmentation of o_i.

Given N such quadruples $(o_i, o_i', h_i^a, h_i^m)$, we construct positive and contrasting negative pairs as shown in Figure 3. In L_{temporal}, for each o_i, o_i' is positive and all other objects o_js and o_j's are negatives. In L_{hand}, for each o_i, $[h_i^a, h_i^m]$ (hand appearance and motion) serves as the positive while all other objects o_j's and hands $[h_j^a, h_j^m]$s are negatives; and for each $[h_i^a, h_i^m]$, o_i is positive and all other objects o_js and hands $[h_j^a, h_j^m]$s are negatives. All crops o_i, o_i', h_i^a are transformed using the standard SimCLR augmentations.

We setup contrastive losses by passing object and hand crops through convolutional trunks ϕ_o and ϕ_h, respectively. We use a projection head f_o for L_{temporal}, and 2 projection heads f_h, g_h (for object and hand crops, respectively) for L_{hand}. h_i^m is encoded via positional encoding and appended to $\phi_h(h_i^a)$ before being fed into projection head g_h. We use cosine similarity, and the normalized temperature scaled cross-entropy loss (*NT-Xent*) following SimCLR [6].

We call our full formulation with both these loss terms as Temporal SimCLR with Object-Hand Consistency or TSC+OHC. We also experiment with Temporal SimCLR or TSC that only uses the temporal term (*i.e*. setting λ to 0). The output of these formulations is ϕ_o, which is our state-sensitive feature representation. In Section 4.1, we evaluate the quality of ϕ_o on an object state classification task.

3.2. Object Affordances via Context Prediction

The next aspect of interactive object understanding that we tackle is to infer what interactions do objects placed in scenes afford and where, which we refer to jointly as object

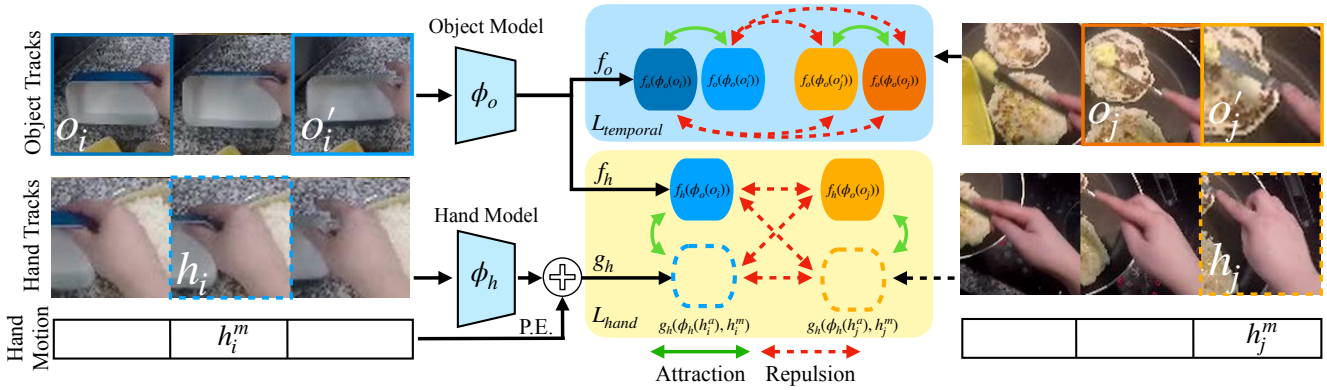

Figure 3. Temporal SimCLR with Object-Hand Consistency (TSC+OHC). Given batches of quadruples containing object crops pairs o_i, o_i', alongside the corresponding hand crop h_i^a and hand motion h_i^m, TSC+OHC employs two losses L_{temporal} and L_{hand}. L_{temporal} encourages object crops close in time to be close to one another, while being far away from other object crops. L_{hand} encourages corresponding object and hands to be close to one another, while being far away from other objects and hands. Different encoders are used for objects and hands (ϕ_o and ϕ_h), and different heads (f_o and f_h) are used for objects for L_{temporal} and L_{hand}. Best viewed in color.

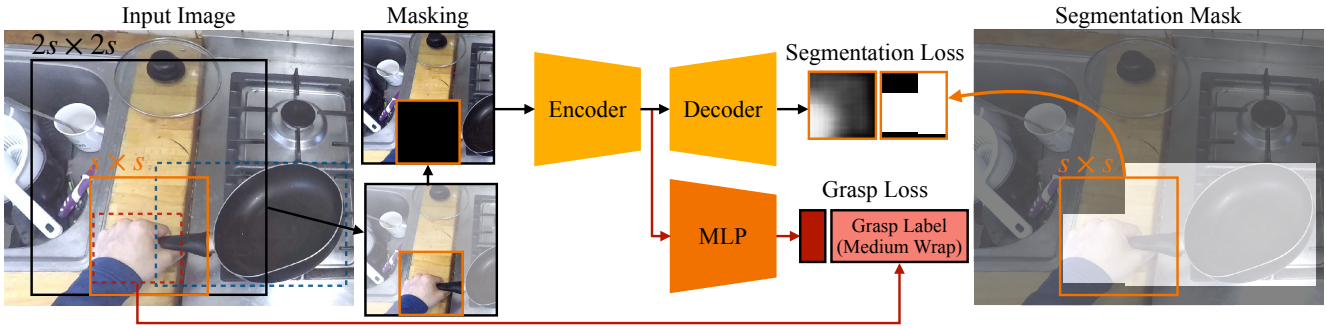

Figure 4. Affordances via Context Prediction (ACP). We sample a patch (orange) from the input image around the detected hand (shown on the left). We then consider a context region (black) of twice the size around the sampled patch containing parts of the object being interacted with. We mask out the sampled patch (Masking) to hide the hand. Our model uses the surrounding context to make predictions for probability of interaction and grasps afforded in the masked region. We paste the hand and object boxes to generate supervision for interaction regions. Supervision for grasp prediction branch comes by running a network trained on GUN71 dataset [51] on the hand crop.

affordances. Specifically, we want to infer a) the Regions of Interaction (ROI) in the scene (*i.e.* pixels that are likely to be interacted with when undertaking some common actions), and b) the hand-grasp type that is applicable at that region.

Information about both these aspects is directly available in egocentric videos. As hands interact with objects, we observe where they touch and via what grasp. However, learning models from such data as-is is hard; wherever we have the hand for supervision, we also have the same hand that trivially reveals the information that we want to predict. As a result, a naively trained model won't learn anything about the underlying object. To circumvent this issue, we propose a context prediction task: prediction of the hand locations and grasp type from image patches around the hand, but with hands *masked out*. Our context prediction task encourages the model to use the context around an object to predict regions of interaction. For instance in Figure 4, the model can predict the region-of-interaction (location of the handle) from part of the pan visible in the context region. We call our model *Affordances via Context Prediction (ACP)*.

Data Generation. Our data generation process, shown in Figure 4, assumes detections for hands, object-of-interaction, contact state, and grasp-type (see Figure 2). Starting with the hand that are in contact state, we sample a $s \times s$ patch around the detected hand. We crop out a $2s \times 2s$ *asymmetric* context region around this patch, with the $s \times s$ hand patch being at the bottom center of this context region. We mask out the $s \times s$ hand patch to obtain a masked context region that serves as the input to our model. The goal for the model is to predict a) the segmentation mask for the hand (and optionally also the object-of-interaction) inside the masked region, and b) the grasp-type exhibited by the hand. Supervision for these comes from the detections and the grasp predictions as described above. As the detector from [55] only outputs boxes, we derive an approximate segmentation mask by pasting the detection boxes. We also sample additional positives from around the object-of-interaction detections and negatives from the remaining image. We sample patches at varying scale and reshape them to 128×128 before feeding them into our network.

Model Architecture and Training. The masked context region is processed through a ResNet-50 encoder, followed by two separate heads to predict the segmentation mask and the grasp type. The segmentation head uses a deconvolutional decoder to produce 64×64 segmentation masks, and is trained using binary cross-entropy loss with the positive class weighed by a factor of 4. The grasp-type prediction uses 2 fully-connected layers to predict the applicable grasp types. As more than one grasp is applicable, we model it as a multi-label problem and train using independent binary cross-entropy losses for each grasp-type. For each example, the highest scoring class from GUN71 model is treated as positive, lowest 15 are treated as negatives, and the remaining are not used for computing loss.

Inference. For inference, we sample patches densely at 3 different scales. We reshape them to 128×128 and mask out the 64×64 bottom center region, before feeding them into our model. Predictions from the patches are pasted back onto the original image to generate per-pixel probability for a) interaction, and b) afforded hand grasps.

Though we only considered predicting coarse segmentation and grasp-types our contextual prediction framework is more general. Given appropriate pre-trained models, ACP can be trained for richer hand representations such as fine-grained segmentation, 2D or 3D hand pose.

4. Experiments

We train our models on in-the-wild videos from EPIC-KITCHENS [10]. Our experiments test the different aspects of interactive object understanding that we pursue: state-sensitive features (Section 4.1), and object affordance prediction (*i.e*. identifying regions-of-interaction (Section 4.2) and predicting hand grasps afforded by objects (Section 4.3)). We focus on comparing different sources of supervision, and on evaluating our design choices. As we pursue relatively new tasks, we collect two labeled datasets on top of EPIC-KITCHENS to support the evaluation: EPIC-STATES for state-sensitive feature learning and EPIC-ROI for regions-of-interaction. We adapt the YCB-Affordance benchmark [9] for afforded hand-grasp prediction.

All our experiments are conducted in the challenging setting where there is *no overlap between training and testing participants* for EPIC-KITCHENS experiments,[1] and *no overlap in objects* for experiments on YCB-Affordance.

4.1. State Sensitive Features for Objects

We measure the state sensitivity of our learned feature space ϕ_o, by testing its performance for fine-grained object state classification. We design experiments to measure the effectiveness of focusing on the hands to derive a) data and b) supervision for learning; and our choice of learning method. We also compare the quality of our self-supervised features to existing methods for learning such features via: action classification on EPIC-KITCHENS and state classification on Internet data [24].

Object State Classification Task and Dataset. For evaluation, we design and collect EPIC-STATES, a labeled object state classification dataset. EPIC-STATES builds upon the raw data in the EPIC-KITCHENS dataset and consists of 10 state categories: OPEN, CLOSE, INHAND, OUTOFHAND, WHOLE, CUT, RAW, COOKED, PEELED, UNPEELED. We selected these state categories as they are defined somewhat unambiguously and had enough examples in the EPIC-KITCHENS dataset. EPIC-STATES consists of 14,346 object bounding boxes from the EPIC-KITCHENS dataset (2018 version), each labeled with 10 binary labels corresponding to the 10 state classes. We split the dataset into training, validation, and testing sets based on the participants, *i.e*. boxes from same participant are in the same split.

To maximally isolate impact of pre-training, we only train a linear classifier on representations learned by the different methods. We report the mean average precision across these 10 binary state classification tasks. We also consider two settings to further test generalization: a) low training data (only using 12.5% of the EPIC-STATES train set), and b) testing on novel object categories (by holding out objects from EPIC-STATES train set).

Implementation Details. *Object-of-Interaction Tracks.* We construct tracks by linking together hand-associated object detections with IoU $\geq$ 0.4 in temporally adjacent frames. We median filter the object box sizes to minimize jumps due to inaccurate detections. This resulted in 61K object tracks (on average 2.2s long) for training. We extract patches at 10 fps from these tracks.

Model Architecture. All models use the ResNet 18 [23] backbone initialized with ImageNet pre-training. We average pooled the 4×4 output from the ResNet 18 backbone and introduced 2 fully connected layers to arrive at a 512 dimensional embedding for all models.

Self-supervision Hyper-parameters. Our proposed models (TSC, TSC+OHC) use standard data augmentations: color jitter, grayscale, resized crop, horizontal flip, and Gaussian blur. Temporal augmentation frames o'_i were within one fourth of the track length. For the TSC+OHC model: hand boxes within 0.3s from the object boxes were considered as corresponding and h^m_i was computed using 3 consecutive frames. See other details in Supplementary.

Results. Table 1 reports the mean average precision (higher is better) for object state classification on the EPIC-STATES test set. We also report the standard deviation across 3 pre-training runs. We compare among our models and against

[1] Note that the detector from [55] was trained on 18K labeled frames from the EPIC-KITCHENS dataset. To ensure that our trainings only see realistic predictions, we use *leave one out* predictions from [55]: we split the train set into 5 parts by participants, retrain [55] on 4, use predictions on the 5$^{\text{th}}$ (*i.e*. unseen participants); and repeat this 5 times over.

Table 1. Mean average precision for object state classification on the EPIC-STATES test set ($\mu \pm \sigma$ over 3 pre-training seeds). Our self-supervised features outperform features from ImageNet-pretraining, other self-supervision (TCN, SimCLR), and even semantic supervision across all settings. Performance boost is larger in harder settings: low-data and generalization to novel objects.

	Novel Objects		All Objects	
Linear classifier training data	12.5%	100%	12.5%	100%
ImageNet Pre-trained	70.2 ±0.0	74.5 ±0.0	78.2 ±0.0	83.1 ±0.0
TCN [53]	56.1 ±1.9	63.9 ±1.1	62.5 ±0.8	73.4 ±1.4
SimCLR [6]	71.9 ±0.2	77.1 ±1.0	77.4 ±1.0	81.0 ±0.9
SimCLR + TCN	63.7 ±0.3	68.4 ±1.6	72.9 ±1.3	77.4 ±1.2
Semantic supervision				
via EPIC action classification	70.9 ±1.9	77.0 ±0.9	72.1 ±0.8	77.9 ±1.3
via MIT States dataset [24]	70.1 ±1.4	73.9 ±0.8	76.4 ±0.6	81.5 ±1.3
Ours [TSC]	74.5 ±0.9	80.2 ±0.4	81.4 ±1.0	84.2 ±1.0
Ours [TSC+OHC]	**79.7 ±0.6**	**81.8 ±0.4**	**82.6 ±0.2**	**84.8 ±0.4**

Figure 5. Object in similar states. Nearest neighbors in our learned feature space exhibit similar state.

a) ImageNet pre-training (*i.e*. no further self-supervised pre-training), b) non-temporal self-supervision via SimCLR [6], c) an alternate temporal self-supervision method (Time Contrastive Networks, TCN [53]), and d) semantic supervision from action classification on EPIC-KITCHENS and state classification on MIT States dataset [24]. We describe these comparison points as we discuss our key takeaways.

Features from TSC and TSC+OHC are more state-sensitive than ImageNet features. ImageNet pre-trained features provide a strong baseline with an mAP of 83.1%. TSC and TSC+OHC boost performance to 84.2% and 84.8%, respectively. Improvements get amplified in the challenging low-data and novel category settings for all models, with our full model TSC+OHC improving upon ImageNet features by 4.4% and 9.5%, respectively. These trends are also borne out when we visualize nearest neighbors in the learned feature spaces in Figure 5.

TSC and TSC+OHC outperform other competing self-supervision schemes. Temporal SimCLR, even by itself, is more effective than vanilla SimCLR that has access to the same crops but ignores the temporal information. We also outperform TCN [53], a leading method for temporal self-supervision, and TCN combined with SimCLR. TCN uses negatives from the same track. These are harder to identify in EPIC-KITCHENS because of the large variability in time-scales at which changes occur (*e.g*. OPEN *vs*. CHOP action).

Supervision from object-hand consistency improves performance. TSC+OHC improves over just TSC by 0.6% with larger gains (of up to 5.2%) in the more challenging

Figure 6. Objects affording similar hands. We retrieve objects that are associated with hands having features similar to the query hand. Objects that are being interacted with similarly get retrieved.

novel category and limited data settings. This confirms our hypothesis that observation of what hands are doing, aids the understanding of object states. Figure 6 shows some nearest neighbors retrievals that further support this.

TSC and TSC+OHC models outperform semantically supervised models. Conventional wisdom would have suggested pre-training a model on images gathered from the Internet for this or related tasks. MIT States dataset from Isola *et al*. [24] is the largest such dataset with 32,915 training images labeled with applicable adjectives. Surprisingly, our self-supervised models outperform features learned through supervised training on this dataset by 3.3 to 9.6%, perhaps due to the domain gap between Internet and egocentric data.

Another common belief is to equate action classification to video understanding. We assess this by comparing against features from the action classification task on EPIC-KITCHENS. This model was trained on our tracks using the most common 32 action labels along with their temporal extent, available as part of the EPIC-KITCHENS dataset. Both TSC and TSC+OHC features outperform action classification features by 3 to 10%. Thus, while the action classification task is useful for many applications, it fails to learn good state-sensitive features.

Ablations. In Supplementary, we compare alternate ways of obtaining tracks when learning with TSC. We ablate two aspects: what we track (background crop, background object, object-of-interaction) and how we track it (no tracking, off-the-shelf tracker [35], hand-context). Ablations reveal the utility of object-of-interaction tracks particularly as they enable use of hand consistency. We also study the role of appearance and motion individually for representing the hand. We found both to be useful over TSC with motion being more important than appearance.

4.2. Regions of Interaction

Regions-of-Interaction Task and Dataset. We design and collect EPIC-ROI, a labeled region-of-interaction dataset. EPIC-ROI builds on top of the EPIC-KITCHENS dataset, and consists of 103 diverse images with pixel-level annotations for regions where human hands *frequently* touch in everyday interaction. Specifically, image regions that afford any of the most frequent actions: TAKE, OPEN, CLOSE, PRESS, DRY, TURN, PEEL are considered as positives. We manually watched video for multiple participants to define a) object categories, and b) specific regions within each category where participants interacted while conducting any

Figure 7. Images from the proposed EPIC-ROI dataset. Each image is annotated for regions of interaction *i.e*. where the human participants frequently interact with. Every annotation is also labeled with one of four attributes: COCO objects, Non-COCO objects, COCO parts, or Non-COCO parts.

of the 7 selected actions. These 103 images were sampled from across 9 different kitchens (7 to 15 images with minimal overlap, from each kitchen). EPIC-ROI is only used for evaluation, and contains 32 val images and 71 test images. Images from the same kitchen are in the same split. The Regions-of-Interaction task is to score each pixel in the image with the probability of a hand interacting with it. Performance is measured using average precision.

To enable detailed analysis, each annotated region is assigned two binary attributes: a) Is-COCO-object (if region is on an object that is included in the COCO dataset), b) Is-whole-object (if region covers the whole object). This results in 4 sub-classes (see Figure 7), allowing evaluation on more challenging aspects: *e.g*. small objects that are not typically represented in object detection datasets such as knobs (Non-COCO Object), or when interaction is localized to a specific object part such as the pan-handle (COCO Part) or the cutting-board-edges (Non-COCO Part). We also evaluate in the 1% SLACK setting where regions within 20 pixels (1% of image width) of the segmentation boundaries is ignored to discount small leakage in predictions.

Implementation Details. We train our model on 250 videos from the 2018 EPIC-KITCHENS dataset. We exclude videos from the 9 kitchens used for evaluation in EPIC-ROI. Details of the grasp classification branch are in Section 4.3.

Results. Table 2 reports the average precision. We compare to three classes of methods: a) objectness based approaches SalGAN [46] and DeepGaze2 [30], trained using human gaze data / manual labels; b) instance segmentation based approaches that use Mask RCNN [22] predicted masks for all / relevant classes; and c) interaction hotspots from Nagarajan *et al*. [42] that derives supervision from manually annotated object bounding boxes and action labels in the EPIC-KITCHENS dataset. Given the strong performance of Mask RCNN-based methods, we also report the performance by aggregating predictions from Mask RCNN with ACP and next most competitive baseline, DeepGaze2. Aggregation is done using a weighted summation of predictions (weight selected using validation performance).

Overall, Mask-RCNN when restricted to relevant categories, performs the best. This is not surprising as it is supervised using over 1 million object segmentation masks. However, its performance suffers on non-COCO objects or their parts. Methods that utilize more general supervision start to do better. And in spite of not being trained on any segmentation masks at all, ACP (*Ours*) is able to outperform past methods. It starts to approach the performance of Mask RCNN, particularly in the 1% SLACK setting.

When combined with Mask RCNN, ACP achieves the strongest performance across all categories. It improves upon the Mask RCNN based method by 4.7%, indicating that our method is able to effectively learn about objects not typically included in detection datasets (*e.g*. stove knobs), and object parts (*e.g*. handle for fridges and drawers). Furthermore, our method provides a more complete interactive understanding by also predicting afforded grasps as we discuss in Section 4.3 and show in Figure 8.

Ablations. Experiments in Supplementary study the effects of variations in network input (not hiding hands, symmetric context, not filtering based on contact), model architecture, and data sampling and supervision (using just objects, or using just hands, or using hand masks rather than boxes). We find that all design choices contribute to ACP's performance. Further improvements can be had from richer hand understanding (segmentation masks *vs*. box masks).

4.3. Hand Grasps Afforded by Objects

Grasps Afforded by Objects (GAO) Task and Dataset. We use the YCB-Affordance dataset [9] to evaluate performance at the Grasps Afforded by Objects (GAO) task. The dataset annotates objects in the scenes from [62] with all applicable grasps from a 33-class taxonomy [15]. We split the dataset into training (110K images, 776K grasps, used only to obtain a supervised ceiling), validation (60 images, 230 grasps) and testing (180 images, 760 grasps). Val and test sets contain *novel* objects not present in the training set. Given an image with a segmentation mask for the object under consideration, the GAO task is to predict the grasps afforded by the object. As multiple grasps are applicable to each object, we measure AP for each grasp independently and report mAP across the 7 (of 33) grasps present in the val and test sets.

Implementation Details. The grasp prediction branch in ACP is trained on predictions from a grasp classification model trained on the GUN71 dataset. We only use the 33 classes relevant to the task on the YCB-Affordance dataset from the 71-way output. To test our grasp affordance prediction on YCB-Affordance objects, we average the spatial grasp scores over the pixels belonging to the object mask.

We found it useful to adapt the GUN71 classifier to EPIC-KITCHENS to generate good supervision. This was done via self-supervision by using an additional L_{temporal} loss on the EPIC-KITCHENS *hand* tracks (analogous to one used for objects in Section 3.1) while training on GUN71.

Table 2. Average Precision for Region-of-Interaction Prediction. We report the overall AP and AP across the different types of interaction regions. We also report AP with 1% SLACK at the boundaries where we don't penalize any leakage at regions within 20 pixels (1% of image width) of the mask boundaries. Without training on segmentation masks, our method outperforms methods based on objectness (SalGAN and DeepGaze2), action classification (Interaction Hotspots), and are able to come close to Mask RCNN that is trained with supervised segmentation masks. We achieve the strongest performance across all categories when combined with Mask RCNN. Highest numbers are boldfaced and the second highest are italicized.

	Overall		**COCO Objects**		**Non-COCO Objects**		**COCO Parts**		**Non-COCO Parts**	
Slack at segment boundaries	0%	1%	0%	1%	0%	1%	0%	1%	0%	1%
Mask RCNN [all]	41.9	46.7	40.6	45.0	13.3	16.0	11.5	14.4	3.6	4.3
Mask RCNN [relevant]	**64.0**	**70.0**	**72.2**	**78.1**	22.8	28.7	**31.0**	**39.6**	6.8	9.1
Interaction Hotspots [42]	43.8	52.0	26.5	33.9	23.0	29.5	12.2	16.7	6.9	9.6
SalGAN [46]	48.7	56.4	40.8	49.1	24.5	31.0	11.4	15.7	4.7	6.4
DeepGaze2 [30]	55.7	64.6	44.8	55.1	**35.8**	**45.4**	11.4	16.5	**7.4**	*10.8*
ACP (*Ours*)	*57.4*	*67.3*	*49.6*	*60.8*	*33.7*	*44.7*	*14.7*	*22.5*	*7.2*	**11.3**
Mask RCNN+DeepGaze2	*66.6*	*72.9*	*74.4*	*80.1*	*26.2*	*33.5*	*31.7*	*40.6*	*7.1*	*9.8*
Mask RCNN+ACP (*Ours*)	**68.7**	**76.4**	**76.2**	**83.0**	**31.1**	**41.9**	**32.5**	**43.7**	**7.4**	**11.4**

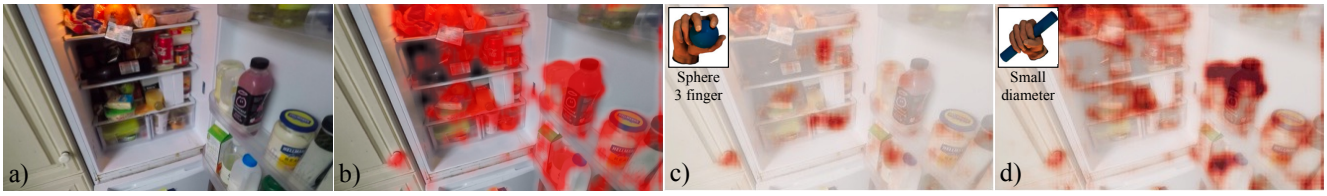

Figure 8. Object Affordance Predictions. For the input image (shown in (a)), we show the predicted regions of interaction in (b). Our method successfully detects multiple possible regions of interaction: bottles, jars, general objects in the fridge, and door knobs. We also visualize the per-pixel probability of affording the sphere 3 finger grasp in (c), and small diameter grasp in (d). The sphere 3 finger grasp is predicted for the door knob, bottle caps and cans; while the small diameter grasp is predicted for bottles, jars and cans. Thumbnails visualizing hand grasps reproduced from [15].

Results. For reference, chance performance is 30.2%, and supervised performance is 56.8%. The supervised method is trained on YCB-Affordance using ground truth annotations for afforded hand grasps on the 110K training images for the 15 training objects. Our method achieves an mAP of 38.1%. It reduces the gap between chance performance and the supervised method by 30%. Adaptation using L_{temporal} on hands helped (34.3% *vs*. 38.1%).

5. Discussion and Limitations

We have shown that observation and analysis of humans hands interacting with the environment is a rich source of information for learning about objects and how to interact with them; even when using a relatively crude understanding of the hand via 2D boxes. A richer understanding of hands (through segmentation, fine-grained 2D / 3D pose, and even 3D reconstruction) would enable a richer understanding of objects in the future. Our work relies on off-the-shelf models for generating data and supervision, and is limited by the quality of their output.

The ACP model in Section 3.2, doesn't look at the pixels it is making predictions on. This causes our predictions to not be as well-localized. Our EPIC-ROI task requires fine-grained reasoning for large objects (*e.g*. microwaves), but not as much for small objects because of subjectivity in annotation. Collecting fine-grained datasets for where we can interact with small objects in scenes will enable better evaluation. Similarly, large-scale in-the-wild datasets for evaluation of grasps afforded by objects can help. Finally, we tackled different aspects of interactive object understanding in isolation, a joint formulation could do better.

Ethical considerations, bias, and potential negative societal impact: Egocentric data is of sensitive nature. We relied on existing public data from EPIC-KITCHENS dataset (which obtained necessary consent and adopted best practices). Though, our self-supervised techniques mitigate bias introduced during annotation, we acknowledge that our models inherit and suffer from bias (*e.g*. what objects are present, their appearance and usage) present in the raw videos in EPIC-KITCHENS dataset. As with all AI research, we acknowledge potential for negative societal impact. Interactive object understanding can enable many useful applications (*e.g*. building assistive systems), but could also be used for large-scale automation which, if not thought carefully about, could have negative implications.

Acknowledgements: We thank David Forsyth, Anand Bhattad, and Shaowei Liu for useful discussion. We also thank Ashish Kumar and Aditya Prakash for feedback on the paper draft. This material is based upon work supported by NSF (IIS-2007035), DARPA (Machine Common Sense program), and an Amazon Research Award.

References

[1] Andrea Bandini and José Zariffa. Analysis of the hands in egocentric vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8709–8719, 2019. 2

[3] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems (RSS)*, volume 3. Ann Arbor, Michigan;, 2016. 2

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 2, 3, 6

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[9] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Gregory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5, 7

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 3, 5

[11] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *Computer Vision and Image Understanding*, 149:98–112, 2016. 2

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 2

[13] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 224–233, 2020. 2

[14] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[15] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015. 2, 7, 8

[16] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 732–745. Springer, 2012. 2

[17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 409–419, 2018. 2

[18] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3497–3506, 2019. 2

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[20] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 671–678, 2010. 2

[21] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 2

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2969, 2017. 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[24] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391, 2015. 2, 5, 6

[25] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[26] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 2

[27] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–214. Springer, 2012. 2

[28] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 5:3352–3359, 2020. 2

[29] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 831–847. Springer, 2014. 2

[30] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7, 8

[31] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021. 2

[32] Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15772–15781, 2021. 2

[33] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015. 2

[34] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018. 2

[35] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[36] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3706–3715, 2020. 2

[37] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*, 2017. 2

[38] Priyanka Mandikal and Kristen Grauman. Dexterous robotic grasping with object-centric visual affordances. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 1, 2

[39] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2021. 1

[40] Kaichun Mo, Leonidas Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[41] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[42] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8688–8697, 2019. 2, 7, 8

[43] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[44] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. EGO-TOPO: Environment affordances from egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 163–172, 2020. 2

[45] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9890–9900, 2020. 2

[46] Junting Pan, Cristian Canton-Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i-Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *CoRR*, abs/1701.01081, 2017. 7, 8

[47] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[48] Vladimír Petrík, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Learning object manipulation skills via approximate state estimation from real videos. *Proceedings of the Conference on Robot Learning (CoRL)*, 2020. 2

[49] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413, 2016. 2

[50] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3696–3705, 2017. 2

[51] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3889–3897, 2015. 2, 3, 4

[52] Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2020. 1

[53] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141, 2018. 2, 6

[54] Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. In *Robotics: Science and Systems (RSS)*, 2017. 1

[55] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 5

[56] Dandan Shan, Richard E.L. Higgins, and David F. Fouhey. COHESIV: Contrastive object and hand embedding segmentation in video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[57] Jin Sun, Hadar Averbuch-Elor, Qianqian Wang, and Noah Snavely. Hidden footprints: Learning contextual walkability from 3d human trails. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 192–207. Springer, 2020. 2

[58] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2019. 2

[59] Spyridon Thermos, Gerasimos Potamianos, and Petros Daras. Joint object affordance reasoning and segmentation in rgb-d videos. *IEEE Access*, 9:89699–89713, 2021. 2

[60] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with novel objects via zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[61] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2596–2605, 2017. 2

[62] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 7

[63] Yezhou Yang, Cornelia Fermuller, Yi Li, and Yiannis Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 400–408, 2015. 2

[64] Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2021. 1

[65] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1058–1067, 2017. 2