Grounded Human-Object Interaction Hotspots From Video

Tushar Nagarajan[*]
UT Austin
tushar@cs.utexas.edu

Christoph Feichtenhofer
Facebook AI Research
feichtenhofer@fb.com

Kristen Grauman
UT Austin and Facebook AI Research
grauman@fb.com

Abstract

Learning how to interact with objects is an important step towards embodied visual intelligence, but existing techniques suffer from heavy supervision or sensing requirements. We propose an approach to learn human-object interaction "hotspots" directly from video. Rather than treat affordances as a manually supervised semantic segmentation task, our approach learns about interactions by watching videos of real human behavior and anticipating afforded actions. Given a novel image or video, our model infers a spatial hotspot map indicating where an object would be manipulated in a potential interaction—even if the object is currently at rest. Through results with both first and third person video, we show the value of grounding affordances in real human-object interactions. Not only are our weakly supervised hotspots competitive with strongly supervised affordance methods, but they can also anticipate object interaction for novel object categories. Project page: http://vision.cs.utexas.edu/projects/interaction-hotspots/

1. Introduction

Today's visual recognition systems know how objects *look*, but not how they *work*. Understanding how objects function is fundamental to moving beyond passive perceptual systems (*e.g.*, those trained for image recognition) to active, embodied agents that are capable of both perceiving and interacting with their environment—whether to clear debris in a search and rescue operation, cook a meal in the kitchen, or even engage in a social event with people. Gibson's theory of affordances [17] provides a way to reason about object function. It suggests that objects have "action possibilities" (*e.g.*, a chair affords sitting, a broom affords cleaning), and has been studied extensively in computer vision and robotics in the context of action, scene, and object understanding [22].

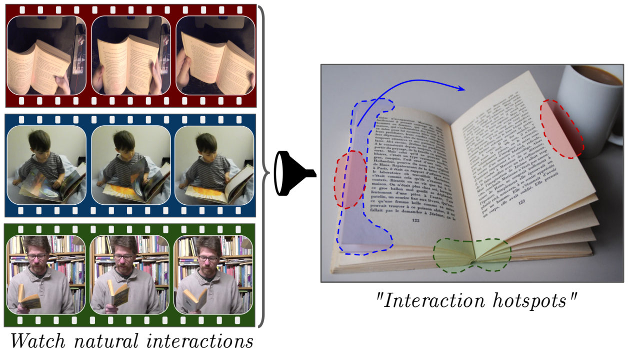

Figure 1: Envisioned concept. We propose to learn object affordances directly from videos of people naturally interacting with objects. The resulting representation of "interaction hotspots" is grounded in real human behavior from video, rather than manual image annotations. See Sec. 4 for examples on video datasets.

However, the abstract notion of "what actions are possible?" addressed by current affordance learning methods is only half the story. For example, for an agent tasked with sweeping the floor with a broom, knowing that the broom handle *affords holding* and the broom *affords sweeping* is not enough. The agent also needs to know the best way to grasp the object, the specific points on the object that need to be manipulated for a successful interaction, how the object is used to achieve a goal, and even what it suggests about how to interact with *other* objects.

Learning how to interact with objects is challenging. Traditional methods face two key limitations. First, methods that consider affordances as properties of an object's shape or appearance [36, 18, 24] fall short of modeling actual object use and manipulation. In particular, learning to segment specified object parts [37, 48, 36, 38] can capture annotators' expectations of what is important, but is detached from real interactions, which are dynamic, multimodal, and may only partially overlap with part regions (see Figure 1). Secondly, existing methods are limited by their heavy supervision and/or sensor requirements. They assume access to training images with manually drawn masks or keypoints [45, 10, 12] and some leverage additional sensors like depth [31, 65, 66] or force gloves [3], all of which restrict scalability. Such bottlenecks also deter generaliza-

[*] Work done during internship at Facebook AI Research.

tion: exemplars are often captured in artificial lab tabletop environments [36, 31, 48] and labeling cost naturally restricts the scope to a narrow set of objects.

In light of these issues, we propose to learn affordances that are *grounded* in real human behavior directly from videos of people naturally interacting with objects, without any keypoint or mask supervision. Specifically, we introduce an approach to infer an object's *interaction hotspots*—the spatial regions most relevant to human-object interactions. Interaction hotspots link images of *inactive* objects at rest to the actions they afford and where they afford them. By learning hotspots directly from video, we sidestep issues stemming from manual annotations, avoid imposing part labels detached from real interactions, and discover exactly how people interact with objects in the wild.

Our approach works as follows. First, we use videos of people performing everyday activities to learn an action recognition model that can recognize the array of afforded actions when they are *actively in progress* in novel videos. Then, we introduce an anticipation model to distill the information from the video model, such that it can estimate how a static image of an *inactive* object transforms during an interaction. In this way, we learn to anticipate the plausible interactions for an object at rest (*e.g.*, perceiving "cuttable" on the carrot, despite no hand or knife being in view). Finally, we propose an activation mapping technique tailored for fine-grained object interactions to derive interaction hotspots from the anticipation model. Thus, given a new image, we can hypothesize interaction hotspots for an object, even if it is not being actively manipulated.

We validate our model on two diverse video datasets: OPRA [12] and EPIC-Kitchens [7], spanning hundreds of object and action categories, with videos from both first and third person viewpoints. Our results show that with just action and object labels as weak supervision for training video clips, our interaction hotspots can predict object affordances more accurately than prior weakly supervised approaches, with relative improvements up to 25%. Furthermore, we show that our hotspot maps can anticipate object function for novel object classes that are never seen during training, and that our model's learned representation encodes functional similarities that go beyond appearance features.

In summary, we make the following contributions:

- We present a framework that integrates action recognition, a novel anticipation module, and feature localization to learn object affordances directly from video, without manually annotated segmentations/keypoints.
- We propose a class activation mapping strategy tailored for fine-grained object interactions that can learn high resolution, localized activation maps.
- Our approach predicts affordances more accurately than prior weakly supervised methods—and even competitively with strongly supervised methods—and can anticipate object interaction for novel object classes unobserved in the training video.

2. Related Work

Visual Affordances. The theory of affordances [17], originally from work in psychology, has been adopted to study several tasks in computer vision [22]. In action understanding, affordances provide context for action anticipation [32, 43, 64] and help learn stronger action recognition models [30]. In scene understanding, they help decide *where* in a scene a particular action can be performed [46, 18, 59, 9], learn scene geometry [21, 15], or understand social situations [5]. In object understanding, affordances help model object function and interaction [52, 61, 66], and have been studied jointly with hand pose/configuration [29, 53, 3] and object motion [19, 20].

The choice of affordance representation varies significantly in these tasks, spanning across human pose, trajectories of objects, sensorimotor grasps, and 3D scene reconstructions. Often, this results in specialized hardware and heavy sensor requirements (*e.g.*, force gloves, depth cameras). We propose to automatically learn appropriate representations for visual affordances directly from RGB video of human-object interactions.

Grounded Affordances. Pixel-level segmentation of object parts [48, 36, 38] is a common affordance representation, for which supervised semantic segmentation frameworks are the typical approach [36, 45, 38, 10]. These segmentations convey high-level information about object function, but rely on manual mask annotations to train—which are not only costly, but can also give an unrealistic view of how objects are actually used. Unlike our approach, such methods are "ungrounded" in the sense that the annotator declares regions of interest on the objects outside of any interaction context.

Representations that are grounded in human behavior have also been explored. In images, human body pose serves as a proxy for object affordance to reveal modes of interaction with musical instruments [61, 62] or likely object interaction regions [4]. Given a video, methods can parse 3D models to estimate physical concepts (velocity, force, etc.) in order to categorize object interactions [65, 66]. For instructional video, methods explore ways to extract object states [1], modes of object interaction [8], interaction regions [12], or the anticipated trajectory of an object given a person's skeleton pose [31].

We introduce a new approach for learning affordance "heatmaps" grounded in human-object interaction, as derived directly from watching real-world videos of people using the objects. Our model differs from other approaches in two main ways. First, no prior about interaction in the form of human pose, hand position, or 3D object reconstruction is used. All information about the interactions is learned directly from video. Second, rather than learn from manually annotated ground truth masks or keypoints [36, 45, 38, 10, 48, 47, 12], our model uses only coarse action labels for video clips to guide learning.

Video anticipation. Predicting future frames in videos has been studied extensively in computer vision [42, 35, 57, 34, 51, 54, 58, 39, 60, 56, 28]. Future prediction has been applied to action anticipation [26, 55, 32, 44], active-object forecasting [16], and to guide demonstration learning in robotics [13, 14, 11]. In contrast to these works, we devise a novel anticipation task—learning object interaction affordances from video. Rather than predict future frames or action labels, our model anticipates correspondences between *inactive* objects (at rest, and not interacted with) and *active* objects (undergoing interaction) in feature space, which we then use to estimate affordances.

3. Approach

Our goal is to learn "interaction hotspots": characteristic object regions that anticipate and explain human-object interactions (see Figure 1). Conventional approaches for learning affordance segmentation only address part of this goal. Their manually annotated segmentations are expensive to obtain, do not capture the dynamics of object interaction, and are based on the annotators' notion of importance, which does not always align with real object interactions. Instead of relying on such segmentations as proxies for interaction, we train our model on a more direct source—videos of people naturally interacting with objects, together with images/frames of these objects at rest. We contend that such videos contain much of the cues necessary to piece together how objects are interacted with.

Our approach consists of three steps. First, we train a video action classifier to recognize each of the afforded actions (Section 3.1). Second, we introduce a novel anticipation model that maps static images of the inactive object to its afforded actions (Section 3.2). Third, we propose an activation mapping technique in the joint model tailored for discovering interaction hotspots on objects, without any keypoint or segmentation supervision (Section 3.3). Given a static image of a novel object, we use the learned model to extract its *hotspot hypotheses* (Section 3.4). Critically, the model can infer hotspots even for object categories unseen during training, and regardless of whether the object is actively being interacted with in the test image.

3.1. Learning Afforded Actions from Video

Our key insight is to learn about object interactions from video. In particular, our approach learns to predict afforded actions across a span of objects, then translates the video cues to static images of an object at rest. In this way, without explicit region labels and without direct estimation of physical contact points, we learn to anticipate object use. Throughout, we use the term "active" to refer to the object when it is involved in an interaction (*i.e*., the status during training) and "inactive" to refer to an object at rest with no interaction (*i.e*., the status during testing).

Let $\mathcal{A}$ denote the set of all afforded actions (*e.g*., *pourable*, *pushable*, *cuttable*), and let $\mathcal{O}$ denote the set of object categories (*e.g*., *pan*, *chair*, *blender*), each of which affords one or more actions in $\mathcal{A}$. During training, we have video clips containing various combinations of afforded actions and objects.

First, we train a video-classification model to predict which afforded action occurs in a video clip. For a video of T frames $\mathcal{V} = \{f_1, ..., f_T\}$ and afforded action class a, we encode each frame using a convolutional neural network backbone to yield $\{x_1, ..., x_T\}$. Each x_t is a tensor with d channels, each with an $n \times n$ spatial extent, with d and n determined by the specific backbone used.[1] These features are then spatially pooled to obtain a d-dimensional vector per frame:

$$g_t = P(x_t) \qquad \text{for } t = 1, \ldots, T, \tag{1}$$

where P denotes the L2-pooling operator. We justify this versus traditional average pooling in Section 3.3.

We further aggregate the frame-level features over time,

$$h_*(\mathcal{V}) = \mathbb{A}(g_1, \ldots, g_T), \tag{2}$$

where $\mathbb{A}$ is a *video aggregation module* that combines the frame features of a video into an aggregate feature h_* for the whole video. In our experiments, we use a long short-term memory (LSTM) recurrent neural network [25] for $\mathbb{A}$. We note that our framework is general and other video classification architectures (*e.g*., 3D ConvNets) can be used.

The aggregate video feature h_* is then fed to a linear classifier to predict the afforded action, which is trained using cross-entropy loss $\mathcal{L}_{cls}(h_*, a)$. Once trained, this model can predict which action classes are observed in a video clip of arbitrary length. See Figure 2 (left) for the architecture.

Note that the classifier's predictions are *object category-agnostic*, since we train it to recognize an afforded action across instances of any object category that affords that action. In other words, the classifier knows $|\mathcal{A}|$ total actions, not $|\mathcal{A}| \times |\mathcal{O}|$; it recognizes *pourable* $+ X$ as one entity, as opposed to *pourable* $+$ *cup* and *pourable* $+$ *bowl* separately. This point is especially relevant once we leverage the model below to generalize hotspots to unfamiliar object classes.

3.2. Anticipation for Inactive Object Affordances

So far, we have a video recognition model that can identify afforded actions in sequences with active human-object

[1] For example, our experiments use a modified ResNet [23] backbone, resulting in $d = 2048$ and $n = 28$, or a $2048 \times 28 \times 28$ feature per frame.

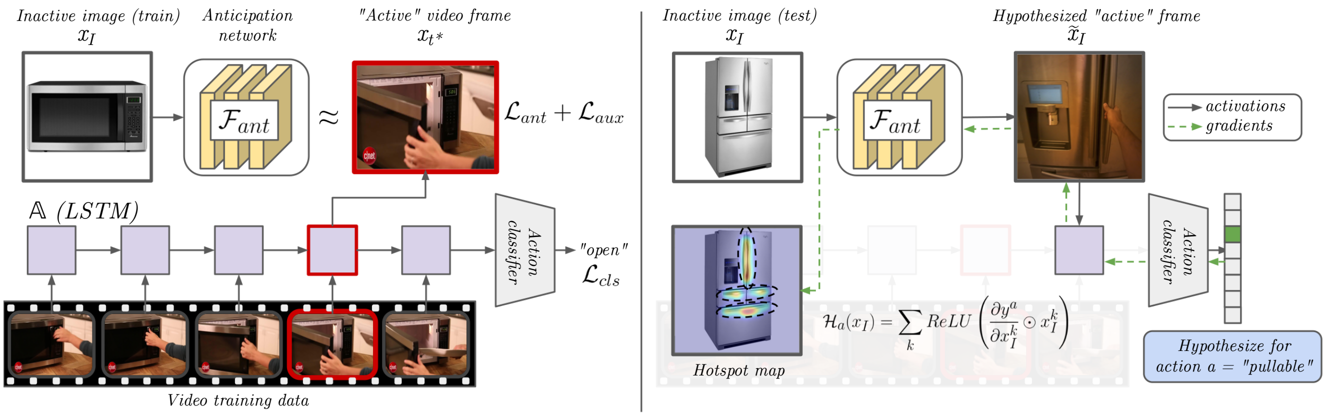

Figure 2: Illustration of our framework for training (left) and testing (right). **Left panel**: The two components of our model —the video action classifier (Sec. 3.1) and the anticipation module with its associated losses (Sec. 3.2 and 3.3)—are jointly trained to predict the action class in a video clip while building an affordance-aware internal representation for objects. **Right panel**: Once trained, our model generates "interaction hotspot" maps for a novel *inactive* object image (top left fridge image). It first hallucinates features that would occur for the object *if it were active* (top right photo), then derives gradient-weighted attention maps over the original image, yielding one map for each action. Our method can infer hotspots even for novel object categories unseen in the training video; for example, learning about opening microwaves helps anticipate how to open the fridge. Note that x_I, $\widetilde{x}_I$ are in feature space, not pixel space.

interactions. This model alone would focus on "active" cues directly related to the action being performed (*e.g.*, hands approaching an object), but would not respond strongly to *inactive* instances—static images of objects that are at rest and not being interacted with. In fact, prior work demonstrates that these two incarnations of objects are visually quite different, to the point of requiring distinct object detectors, *e.g.*, to recognize both open and closed microwaves [41].

We instead aim for our system to learn about object affordances by watching video of people handling objects, then mapping that knowledge to novel inactive object photos/frames. To bridge this gap, we introduce a distillation-based anticipation module $\mathcal{F}_{ant}$ that transforms the embedding of an inactive object x_I, where no interaction is occurring, into its active state where it is being interacted with:

$$\widetilde{x}_I = \mathcal{F}_{ant}(x_I). \tag{3}$$

See Figure 2, top-left. In experiments we consider two sources of inactive object training images x_I: frames from a training sequence showing the object before an action starts (EPIC), or catalog photos of the object shown at rest (OPRA). During training, the anticipation module is guided by the video action classifier, which selects the appropriate *active state* from a given video as the frame x_{t^*} at which the LSTM is maximally confident of the true action:

$$t^* = \underset{t \in 1..T}{\arg\min}\, \mathcal{L}_{cls}(\mathbb{A}(g_1, ..., g_t), a), \tag{4}$$

where a is the true afforded action label, and $\mathcal{L}_{cls}$ is again the cross-entropy loss for classification using the aggregated hidden state at each time t.

We then define a feature matching loss between (a) the anticipated active state for the inactive object and (b) the active state selected by the classifier for the training sequence. This loss requires the anticipation model to hypothesize a grounded representation of what an object would look like during interaction, according to the actual training video:

$$\mathcal{L}_{ant}(x_I, x_{t^*}) = ||P(\widetilde{x}_I) - P(x_{t^*})||_2. \tag{5}$$

Additionally, we include an auxiliary classification loss $\mathcal{L}_{aux}(h_1(\widetilde{x}_I), a)$ to ensure that the single-frame anticipated feature $\widetilde{x}_I$ is predictive of the afforded action, and that our model is robust to processing both single frames *and* video sequences.

Overall, these components allow our model to estimate what a static inactive object may potentially look like—in feature space—if it were to be interacted with. They provide a crucial link between classic action recognition and affordance learning. As we will define next, activation mapping through $\mathcal{F}_{ant}$ then provides information about what spatial locations on the original static image are most strongly correlated to how it would be interacted with.

3.3. Interaction Hotspot Activation Mapping

At test time, given an inactive object image q, our goal is to infer the *interaction hotspot* maps $\mathcal{H}_a$, for all $a \in \mathcal{A}$, each of which is an $H \times W$ matrix summarizing the regions of interest that characterize an object interaction, where H, W denote the height and width of the source image.[2] Intuitively, a hotspot map should pick up on the regions of the

[2] To process a novel video, we simply compute hotspots for each frame.

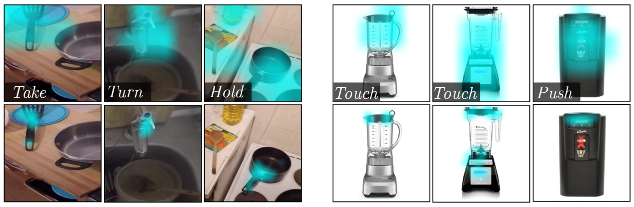

Figure 3: Our method (bottom) vs. traditional action recognition+Grad-CAM (top). Our model generates localized and affordance-relevant activations.

object that would be manipulated or otherwise transform during the action a, indicative of its affordances. Note that there is one map per action $a \in \mathcal{A}$.

We devise an activation mapping approach to go from inactive image embeddings x_I to interaction predictions $\mathcal{F}_{ant}(x_I)$, and finally to hotspots, tailoring it for discovering hotspot maps. For an inactive image embedding x_I and an action class a, we compute the gradient of the score for the action class with respect to each channel of the embedding. These gradients are used to weight individual spatial activations in each channel, acting as an attention mask over them. The positive components of the resulting tensor are accumulated over all channels in the input embedding to give the final hotspot map $\mathcal{H}_a(x_I)$ for the action class:

$$\mathcal{H}_a(x_I) = \sum_k ReLU\left(\frac{\partial y^a}{\partial x_I^k} \odot x_I^k\right), \tag{6}$$

where x_I^k is the k^{th} channel of the input frame embedding and $\odot$ is the element-wise multiplication operator. This is meaningful only when the gradients are not spatially uniform (*e.g.*, not if x_I is average pooled for classification). We use L2-pooling to ensure that spatial locations produce gradients as a function of their activation magnitudes.

Next, we address the spatial resolution. The reduced spatial resolution from repeatedly downsampling features in the typical ResNet backbone is reasonable for classification, but is a bottleneck for learning interaction hotspots. We set the spatial stride of the last two residual stages to 1 (instead of 2), and use a dilation for its filters. This increases the spatial resolution by 4× to n = 28, allowing our heatmaps to capture finer details.

Our technique is related to other feature visualization methods [50, 63, 49]. However, we use a reduced stride and L2 pooling to make sure that the gradients themselves are spatially localized, and like [50], we do not spatially average gradients—we directly weight activations by them and sum over channels. This is in contrast to GradCAM [63, 49] which produces maps that are useful for coarse object localization, but insufficient for interaction hotspots due to their low spatial resolution and diffused global responses. Compared to simply applying GradCAM to an action recognition LSTM (Figure 3, top row), our model produces interaction hotspots that are significantly richer (bottom row). These differences are precisely due to both our anticipation distillation model trained jointly with the recognition model (Sec. 3.2), as well as the activation mapping strategy above. We provide a quantitative comparison in Sec. 4.

3.4. Training and Inference

During training (Figure 2, left), we generate embeddings $\{x_1, ..., x_T\}$ for each frame of a video $\mathcal{V}$. These are passed through $\mathbb{A}$ to generate the video embedding h_*, and then through a classifier to predict the afforded action label a. In parallel, the inactive object image embedding x_I is computed and used to train the anticipation model to predict its *active* state $\widetilde{x}_I$.

The complete loss function for each training instance is:

$$\mathcal{L}(\mathcal{V}, \mathcal{I}, a) = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{ant}\mathcal{L}_{ant} + \lambda_{aux}\mathcal{L}_{aux}, \tag{7}$$

where the λ terms control the weight of each component of the loss, and $\mathcal{I}$ denotes the inactive object image.

For inference on an inactive test image (Figure 2, right), we first generate its image embedding x_I. Then, we hypothesize its *active* interaction embedding $\widetilde{x}_I$, and use it to predict the afforded action scores. Finally, using Equation 6 we generate $|\mathcal{A}|$ heatmaps over x_I, one for each afforded action class. This stack of heatmaps are the *interaction hotspots*. Note that we produce activation maps for the original inactive image x_I, not for the hypothesized active output $\widetilde{x}_I$, *i.e.*, we propagate gradients *through* the anticipation network. Not doing so produces activation maps that are inconsistent with the input image, which hurts performance (see ablation study in Supp).

We stress that interaction hotspots are predictable even for unfamiliar objects. By training the afforded actions across object category boundaries, the system learns the general properties of appearance *and* interaction that characterize affordances. Hence, our approach can anticipate, for example, where an unfamiliar kitchen device might be manipulated, because it has learned how a variety of other objects are interacted with. Similarly, heatmaps can be hallucinated for novel action-object pairs that have not been seen in training (*e.g.*, "cut" using a spatula in Figure 4, bottom row). Please see Supp. for implementation details.

4. Experiments

Our experiments on interaction hotspots explore their ability to describe affordances of objects, to generalize to anticipate affordances of unfamiliar objects, and to encode functional similarities between object classes.

Datasets. We use two datasets:

- **OPRA** [12] contains videos of product reviews of appliances (*e.g.*, refrigerators, coffee machines) collected from YouTube. Each instance is a short video demonstration $\mathcal{V}$ of a product's feature (*e.g.*, pressing a button on a

	Supervision Source	Type	N
EGOGAZE [27]	Recorded eye fixations	Weak	60k
SALIENCY [40, 6, 33]	Manual saliency labels	Weak	10k
OURS	Action, object labels	Weak	20k
IMG2HEATMAP	Manual affordance keypoints	Strong	20k
DEMO2VEC [12]	Manual affordance keypoints, action labels	Strong	20k

Table 1: Supervision source and type for all methods. Our method learns interaction hotspots *without* strong supervision like annotated segmentation/keypoints. N is the number of instances.

coffee machine) paired with a static image $\mathcal{I}$ of the product, an interaction label a (*e.g.*, "pressing"), and a manually created affordance heatmap $\mathcal{M}$ (*e.g.*, highlighting the button on the static image). There are ∼16k training instances of the form $(\mathcal{V}, \mathcal{I}, a, \mathcal{M})$, spanning 7 actions.

- **EPIC-Kitchens** [7] contains unscripted, egocentric videos of activities in a kitchen. Each clip $\mathcal{V}$ is annotated with action and object labels a and o (*e.g.*, cut tomato, open refrigerator) along with a set of bounding boxes $\mathcal{B}$ (one per frame) for objects being interacted with. There are ∼40k training instances of the form $(\mathcal{V}, a, o, \mathcal{B})$, spanning 352 objects and 125 actions. We crowd-source annotations for ground-truth heatmaps $\mathcal{M}$ resulting in 1.8k annotated instances over 20 action and 31 objects (see Supp. for details).

The two video datasets span diverse settings. OPRA has third person videos, where the person and the product being reviewed are clearly visible, and covers a small number of actions and products. EPIC-Kitchens has first-person videos of unscripted kitchen activities and a much larger vocabulary of actions and objects; the person is only partially visible when they manipulate an object. Together, they provide good variety and difficulty to evaluate the robustness of our model.[3] For both datasets, our model uses only the action labels as supervision, and an *inactive* image for our anticipation loss $\mathcal{L}_{ant}$. We stress that (1) the annotated heatmap $\mathcal{M}$ is used *only* for evaluation, and (2) the ground truth is well-aligned with our objective, since annotators were instructed to watch an interaction video clip to decide what regions to annotate for an object's affordances.

While OPRA comes with an image $\mathcal{I}$ of the *exact* product associated with each video instance, EPIC does not. Instead, we crop out inactive objects from frames using the provided bounding boxes $\mathcal{B}$, and randomly select one that matches the object class label in the video. To account for the appearance mismatch, in place of the L2 loss in Equation 5 we use a triplet loss, which uses "negatives" to ensure that inactive objects of the correct class can anticipate active features better than incorrect classes (see Supp. for details).

4.1. Interaction Hotspots as Grounded Affordances

In this section, we evaluate two things: 1) How well does our model learn object affordances? and 2) How well can it infer possible interactions for unfamiliar objects? For this, we train our model on video clips, and generate hotspot maps on *inactive* images where the object is at rest.

Baselines. We evaluate our model against several baselines and state-of-the-art models.

- **Center Bias** produces a fixed Gaussian heatmap at the center of the image. This is a naive baseline to account for a possible center bias [6, 33, 40, 27].
- **LSTM+Grad-CAM** uses an LSTM trained for action recognition with the same action class labels as our method, then applies standard Grad-CAM [49] to get heatmaps. It has no anticipation model.
- **Saliency** is a set of baselines that estimate the most salient regions in an image using models trained directly on saliency annotations/eye fixations: EGOGAZE [27], MLNET [6], DEEPGAZEII [33] and SALGAN [40]. We use the authors' pretrained models.
- **Demo2Vec** [12] is a supervised method that generates an affordance heatmap using context from a video demonstration of the interaction. We use the authors' precomputed heatmap predictions.
- **Img2Heatmap** is a supervised method that uses a fully convolutional encoder-decoder to predict the affordance heatmap for an image. It serves as a simplified version of Demo2Vec that lacks video context during training.

The Saliency baselines capture a generic notion of spatial *importance*. They produce a single heatmap for an image, regardless of action class, and as such, are less expressive than our per-action-affordances. They are *weakly supervised* in that they are trained for a different task, albeit with strong supervision (heatmaps, gaze points) for that task. Demo2Vec and Img2Heatmap are strongly supervised, and represent more traditional affordance learning techniques that learn affordances from manually labeled images [36, 45, 38, 10]. Table 1 summarizes the sources and types of supervision for all methods. Unlike other methods, ours uses only class labels as weak supervision for training.

Grounded Affordance Prediction. First we compare the ground truth heatmaps for each interaction to our hotspots and the baselines' heatmaps. We report error as KL-Divergence, following [12], as well as other metrics (SIM, AUC-J) from the saliency literature [2].

Table 2 (Left) summarizes the results. Our model outperforms all other weakly-supervised methods in all metrics across both datasets. These results highlight that our

[3] Other affordance segmentation datasets [36, 37] have minimal vocabulary overlap with OPRA/EPIC classes, and hence do not permit evaluation for our setting, since we learn from video.

		Grounded affordance prediction						Generalization to novel objects					
		OPRA			EPIC			OPRA			EPIC		
		KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑
	CENTER BIAS	11.132	0.205	0.625	10.660	0.222	0.634	6.281	0.244	0.680	5.910	0.277	0.699
WS	LSTM+GRAD-CAM	8.573	0.209	0.620	6.470	0.257	0.626	5.405	0.259	0.644	4.508	0.255	0.664
	EGOGAZE [27]	2.428	0.245	0.646	2.241	0.273	0.614	2.083	0.278	0.694	1.974	0.298	0.673
	MLNET [6]	4.022	0.284	0.763	6.116	0.318	0.746	2.458	0.316	0.778	3.221	0.361	0.799
	DEEPGAZEII [33]	1.897	0.296	0.720	1.352	0.394	0.751	1.757	0.318	0.742	1.297	0.400	0.793
	SALGAN [40]	2.116	0.309	0.769	1.508	0.395	0.774	1.698	0.337	0.790	1.296	**0.406**	0.808
	OURS	**1.427**	**0.362**	**0.806**	**1.258**	**0.404**	**0.785**	**1.381**	**0.374**	**0.826**	**1.249**	0.405	**0.817**
SS	IMG2HEATMAP	1.473	0.355	0.821	1.400	0.359	0.794	1.431	0.362	0.820	1.466	0.353	0.770
	DEMO2VEC [12]	1.197	0.482	0.847	–	–	–	–	–	–	–	–	–

Table 2: Interaction hotspot prediction results on OPRA and EPIC. **Left:** Our model outperforms other weakly supervised (WS) methods in all metrics, and approaches the performance of strongly supervised (SS) methods *without* the privilege of heatmap annotations during training. **Right:** Not only does our model generalize to new *instances*, but it also accurately infers interaction hotspots for novel object *categories* unseen during training. The proposed hotspots generalize on an object-function level. Values are averaged across three splits of object classes. (↑/↓ indicates higher/lower is better.) DEMO2VEC [12] is available only on OPRA and only for seen classes.

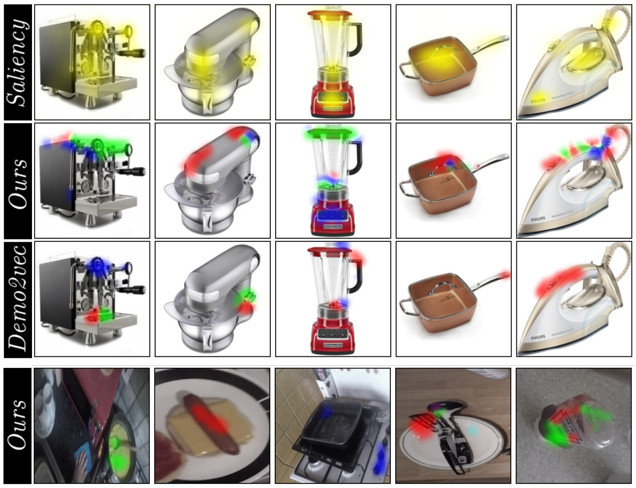

Figure 4: Affordance heatmaps on inactive images. Top: Predicted affordance heatmaps for *hold*, *rotate*, *push* (red, green, blue) on OPRA. **Bottom row:** Predicted heatmaps for *cut*, *mix*, *turn-on* (red, green, blue), on EPIC. Our model highlights spatial affordances consistent with how people interact with the objects. Note that SALIENCY [40] produces only a single "importance" map (yellow). **Last column:** failure cases. Best viewed in color.

model can capture sophisticated interaction cues that describe more specialized notions of importance than saliency.

On OPRA, our model achieves relative improvements of up to 25% (KLD) compared to the strongest baseline, and it matches one of the strongly supervised baseline methods on two metrics. On EPIC, our model achieves relative improvements up to 7% (KLD). EPIC has a much larger, more granular action vocabulary, resulting in fewer and less spatially distinct hotspots. As a result, the baselines that produce redundant heatmaps for all actions artificially benefit on EPIC, though our results remain better.

The baselines have similar trends across datasets. Consistent with the examples in Figure 3, LSTM+GRAD-CAM in Table 2 shows that a simple action recognition model is clearly insufficient to learn affordances. Our anticipation model bridges the (in)active gap between training video and test images, and is crucial for accuracy. All saliency methods perform worse than our model, despite that they may accidentally benefit from the fact that kitchen appliances have interaction regions designed to be visually salient (*e.g.*, buttons, handles). In contrast to our approach, none of the saliency baselines distinguish between affordances; they produce a single heatmap representing "important" salient points. To these methods, the blade of a knife is as important to the action "cutting" as it is to the action "holding", and they are unable to explain objects with multiple affordances. IMG2HEATMAP and DEMO2VEC generate better affordance heatmaps, but at the cost of strong supervision. Our method actually approaches their accuracy without using any manual heatmaps for training.

Please see the Supp. file for an **ablation study** that further examines the contributions of each part of our model. In short, our class activation mapping strategy and propagating gradients all the way through the anticipation model are critical. All elements of the design play a role to achieve our full model's best accuracy.

Figure 4 shows example heatmaps for inactive objects. Our model is able to highlight specific object regions that afford actions (*e.g.*, the knobs on the coffee machine as "rotatable" in column 1) after only watching videos of object interactions. Weakly supervised SALIENCY methods highlight *all* salient object parts in a single map, regardless of the interaction in question. In contrast, our model highlights multiple distinct affordances for an object. To generate comparable heatmaps, DEMO2VEC requires annotated heatmaps for training *and* a set of video demonstrations during inference, whereas our model can hypothesize object functionality without these extra requirements.

Generalization to Novel Objects. Can interaction hotspots

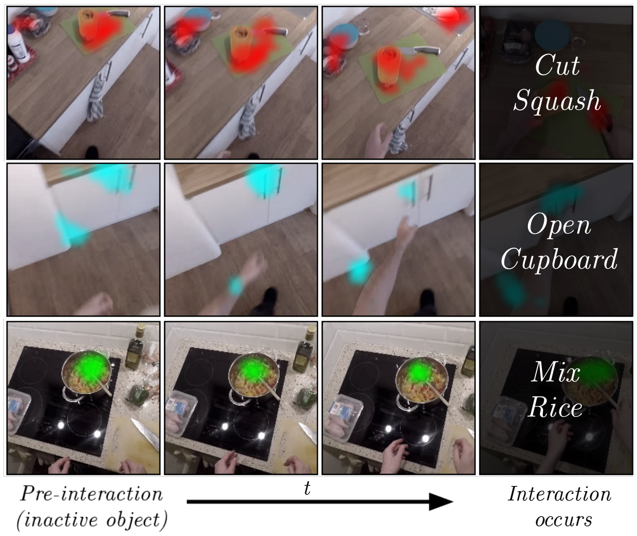

Figure 5: Interaction hotspots on EPIC videos of unseen object classes. Our model anticipates interaction regions for inactive objects at rest (first column), before the interaction happens. Critically, the object categories shown in this figure were *not* seen during training; our model learns to generalize interaction hotspots. For example, there are no cupboards or squashes in the training videos, but our method anticipates how these objects would be opened and cut, respectively. Our method is applied per frame.

infer how *novel* object categories work? We next test if our model learns an *object-agnostic* representation for interaction—one that is not tied to object class. This is a useful property for open-world situations where unfamiliar objects may have to be interacted with to achieve a goal.

We divide the object categories $\mathcal{O}$ into familiar and unfamiliar objects $\mathcal{O} = \mathcal{O}_f \bigcup \mathcal{O}_u$; familiar ones are those seen with interactions in training video and unfamiliar ones are seen only during testing. We leave out 10/31 objects in EPIC and 9/26 objects in OPRA for our experiments, and divide our video train/test sets along these object splits. We train our model only on clips with familiar objects from $\mathcal{O}_f$. While no instances of cupboards, for example, exist in the training split, microwaves and refrigerators do. Instances from these categories are visually distinct, but they are interacted with in very similar ways ("swung open"). If our model can successfully infer the heatmaps for novel, unseen objects, it will show that a general sense of object *function* is learned that is not strongly tied to object *identity*.

Table 2 (Right) shows the results. We see mostly similar trends as the previous section. On OPRA, our model outperforms all baselines in all metrics, and is able to infer the hotspot maps for unfamiliar object categories, despite never seeing them during training. On EPIC, our method remains the best weakly supervised method.

Qualitative results (Figure 5) support our numbers, showing our model applied to video clips from EPIC Kitchens, just before the action occurs. Our model—which was never trained on some objects (*e.g.*, cupboard, squash)—is able to anticipate characteristic spatial locations of interactions *before* the interaction occurs.

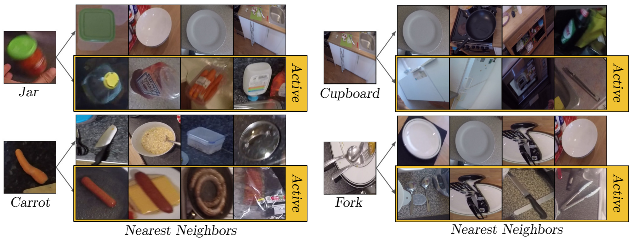

Figure 6: Inactive vs. active object embeddings. By hypothesizing potential interactions with objects, our model learns representations that capture functional similarities between objects across object classes, rather than purely appearance-based similarities.

4.2. Interaction Hotspots for Functional Similarity

Finally, we show how our model encodes functional object similarities in its learned representation for objects. We compare the inactive object embedding space (standard ResNet features) to our predicted active embedding space (output of the anticipation model) by looking at nearest neighbor images in other object classes.

Figure 6 shows examples. Neighbors in the inactive object space (top branch) capture typical appearance-based visual similarities that are useful for object categorization—shapes, backgrounds, etc. In contrast, our active object space (bottom branch, yellow box) reorganizes the objects based on *how* they are interacted with. For example, fridges, cupboards, and microwaves, that are swung open in a characteristic way (top right); knives, spatulas, tongs, that are typically held at their handles (bottom right). Our model learns representations indicative of functional similarity between objects, despite the objects being visually distinct. See Supp. for a clustering visualization on all images.

5. Conclusion

We introduced a method to learn "interaction hotspot" maps—characteristic regions on objects that anticipate and explain object interactions—directly from watching videos of people naturally interacting with objects. Our experiments show that these hotspot maps explain object affordances better than other existing weakly supervised models and can generalize to anticipate affordances of unseen objects. Furthermore, the representation learned by our model goes beyond appearance similarity to encode functional similarity. In future work, we plan to explore how hotspots might aid action anticipation and policy learning for robot-object interaction.

Acknowledgments: Thanks to the authors of Demo2Vec [12] for their help with the OPRA dataset, and to Marcus Rohrbach and Jitendra Malik for helpful discussions. UT Austin is supported in part by ONR PECASE N00014-15-1-2291.

References

[1] Jean-Baptiste Alayrac, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. *ICCV*, 2017. 2

[2] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 2018. 6

[3] Claudio Castellini, Tatiana Tommasi, Nicoletta Noceti, Francesca Odone, and Barbara Caputo. Using object affordances to improve object recognition. *TAMD*, 2011. 1, 2

[4] Chao-Yeh Chen and Kristen Grauman. Subjects and their objects: Localizing interactees for a person-centric view of importance. *IJCV*, 2016. 2

[5] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. *CVPR*, 2018. 2

[6] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *ICPR*, 2016. 6, 7

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. *ECCV*, 2018. 2, 6

[8] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014. 2

[9] Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012. 2

[10] Thanh-Toan Do, Anh Nguyen, Ian Reid, Darwin G Caldwell, and Nikos G Tsagarakis. Affordancenet: An end-to-end deep learning approach for object affordance detection. *ICRA*, 2017. 1, 2, 3, 6

[11] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *CORL*, 2017. 3

[12] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 1, 2, 3, 5, 6, 7, 8

[13] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, 2016. 3

[14] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA*, 2017. 3

[15] 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. *IJCV*, 2014. 2

[16] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *JVCIR*, 2017. 3

[17] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 1979. 1, 2

[18] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *CVPR*, 2011. 1, 2

[19] Abhinav Gupta and Larry S Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 2

[20] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 2009. 2

[21] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 2

[22] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *arXiv preprint arXiv:1807.06775*, 2018. 1, 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[24] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *ICRA: Workshop on Semantic Perception, Mapping, and Exploration*, 2011. 1

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 3

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

[27] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. *ECCV*, 2018. 6, 7

[28] Dinesh Jayaraman, Frederik Ebert, Alexei A Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. *ICLR*, 2019. 3

[29] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *CVIU*, 2011. 2

[30] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJR*, 2013. 2

[31] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *ECCV*, 2014. 1, 2

[32] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *TPAMI*, 2016. 2, 3

[33] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 6, 7

[34] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *ICCV*, 2017. 3

[35] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, 2016. 3

[36] Austin Myers, Ching Lik Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015. 1, 2, 3, 6

[37] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *IROS*, 2016. 1, 6

[38] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 1, 2, 3, 6

[39] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *NeurIPS*, 2015. 3

[40] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 6, 7

[41] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 4

[42] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 3

[43] Nicholas Rhinehart and Kris M Kitani. Learning action maps of large environments via first-person vision. In *CVPR*, 2016. 2

[44] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting from video with online inverse reinforcement learning. *TPAMI*, 2018. 3

[45] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016. 1, 2, 3, 6

[46] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Scenegrok: Inferring action maps in 3d environments. *TOG*, 2014. 2

[47] Johann Sawatzky and Juergen Gall. Adaptive binarization for weakly supervised affordance segmentation. *ICCV: Workshop on Assistive Computer Vision and Robotics*, 2017. 3

[48] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *CVPR*, 2017. 1, 2, 3

[49] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 5, 6

[50] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *ICLR: Workshop*, 2015. 5

[51] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 3

[52] Michael Stark, Philipp Lies, Michael Zillich, Jeremy Wyatt, and Bernt Schiele. Functional object class detection based on learned affordance cues. In *ICVS*, 2008. 2

[53] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. *CVPR*, 2017. 2

[54] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017. 3

[55] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating the future by watching unlabeled video. *CVPR*, 2016. 3

[56] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016. 3

[57] Carl Vondrick and Antonio Torralba. Generating the future with adversarial transformers. In *CVPR*, 2017. 3

[58] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 3

[59] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 2

[60] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NeurIPS*, 2016. 3

[61] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 2

[62] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei. Discovering object functionality. In *ICCV*, 2013. 2

[63] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 5

[64] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, and Qi Tian. Cascaded interactional targeting network for egocentric video analysis. In *CVPR*, 2016. 2

[65] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *CVPR*, 2016. 1, 2

[66] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *CVPR*, 2015. 1, 2