Grounded Human-Object Interaction Hotspots From Video

Tushar Nagarajan
UT Austin

Christoph Feichtenhofer
Facebook AI Research

Kristen Grauman
UT Austin and Facebook AI Research

tushar@cs.utexas.edu
feichtenhofer@fb.com
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].

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 [37, 18, 24] fall short of modeling actual object use and manipulation. In particular, learning to segment specified object parts [38, 49, 37, 39] 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 [46, 10, 12] and some leverage additional sensors like depth [31, 66, 67] or force gloves [3], all of which restrict scalability. Such bottlenecks also deter generaliza-
Our approach predicts affordances more accurately than prior weakly supervised approaches—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, 44, 65] and help learn stronger action recognition models [30]. In scene understanding, they help decide where in a scene a particular action can be performed [47, 18, 60, 9], learn scene geometry [21, 15], or understand social situations [5]. In object understanding, affordances help model object function and interaction [53, 62, 67], and have been studied jointly with hand pose/configuration [29, 54, 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 [49, 37, 39] is a common affordance representation, for which supervised semantic segmentation frameworks are the typical approach [37, 46, 39, 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 [62, 63] 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 [66, 67]. 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 [37, 46, 39, 10, 49, 48, 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 [43, 36, 58, 35, 52, 55, 59, 40, 61, 57, 28]. Future prediction has been applied to action anticipation [26, 56, 32, 45], 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 how objects are interacted with. 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 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 $\mathcal{V} = \{f_1, \ldots, f_T\}$ and afforded action class $a$, we encode each frame using a convolutional neural network backbone to yield $\{x_1, \ldots, x_T\}$. 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.\footnote{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.} These features are then spatially pooled to obtain a $d$-dimensional vector per frame:

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

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,

$$h_{\star}(\mathcal{V}) = \mathcal{A}(g_1, \ldots, g_T),$$

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

The aggregate video feature $h_{\star}$ is then fed to a linear classifier to predict the afforded action, which is trained using cross-entropy loss $L_{\text{cls}}(h_{\star}, a)$. 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
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 [42].

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 $F_{ant}$ 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:

$$\tilde{x}_I = F_{ant}(x_I).$$

(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 $x_{It}$, at which the LSTM is maximally confident of the true action:

$$t^* = \arg\min_{t \in 1..T} \mathcal{L}_{cls}(h_t(g_1, ..., g_t), a),$$

(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:

$$\mathcal{L}_{ant}(x_I, x_{It}) = ||P(\tilde{x}_I) - P(x_{It})||_2.$$  

(5)

Additionally, we include an auxiliary classification loss $\mathcal{L}_{aux}(h_I(x_I), a)$ to ensure that the single-frame anticipated feature $\tilde{x}_I$ 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 $F_{ant}$ 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 $H_a$, 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. Intuitively, a hotspot map should pick up on the regions of the

$$H_a = \sum_L \text{ReLU} \left( \frac{\partial P(x)}{\partial x} \right)_L$$

To process a novel video, we simply compute hotspots for each frame.
We devise an activation mapping approach to go from inactive image embeddings \( x_I \) to interaction predictions \( F_{\text{ant}}(x_I) \), 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:

\[
\mathcal{H}_a(x_I) = \sum_k \text{ReLU} \left( \frac{\partial y^a}{\partial x^k_I} \odot x^k_I \right),
\]

where \( x^k_I \) 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 \( 4 \times \) to \( n = 28 \), allowing our heatmaps to capture finer details.

Our technique is related to other feature visualization methods [51, 64, 50]. However, we use a reduced stride and L2 pooling to make sure that the gradients themselves are spatially localized, and like [51], we do not spatially average gradients—we directly weight activations by them and sum over channels. This is in contrast to GradCAM [64, 50] 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 \( \{x_1, \ldots, x_T\} \) for each frame of a video \( V \). These are passed through \( A \) to generate the video embedding \( h_v \), 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 \( \tilde{x}_I \).

The complete loss function for each training instance is:

\[
\mathcal{L}(V, I, a) = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{ant}} L_{\text{ant}} + \lambda_{\text{aux}} L_{\text{aux}},
\]

where the \( \lambda \) terms control the weight of each component of the loss, and \( I \) 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 \( \tilde{x}_I \), and use it to predict the afforded action scores. Finally, using Equation 6 we generate \( |A| \) 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 \( \tilde{x}_I \), 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 \( V \) of a product’s feature (e.g., pressing a button on a
for our anticipation loss $L$ only the action labels as supervision, and an indicator overlap with OPRA/EPIC classes, and hence do not permit evaluation for our setting, since we learn from video.

...coffee machine) paired with a static image $I$ of the product, an interaction label $a$ (e.g., “pressing”), and a manually created affordance heatmap $M$ (e.g., highlighting the button on the static image). There are $\sim$40k training instances of the form $(V, I, a, M)$, spanning 37 actions.

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

For both datasets, our model uses only the action labels as supervision, and an inactive image for our anticipation loss $L_{ant}$. We stress that (1) the annotated heatmap $M$ 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 $I$ 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 $B$, 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, 34, 41, 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 [50] 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: EOGAZE [27], MLNET [6], DEEPGAZEL [34] and SALGAN [41]. 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’ pre-computed 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 [37, 46, 39, 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

<table>
<thead>
<tr>
<th>Supervision Source</th>
<th>Type</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EGOGAZE</strong> [27]</td>
<td>Recorded eye fixations</td>
<td>Weak</td>
</tr>
<tr>
<td><strong>SALIENCY</strong> [41, 6, 34]</td>
<td>Manual saliency labels</td>
<td>Weak</td>
</tr>
<tr>
<td><strong>OURS</strong></td>
<td>Action, object labels</td>
<td>Weak</td>
</tr>
<tr>
<td><strong>IMG2HEATMAP</strong></td>
<td>Manual affordance keypoints</td>
<td>Strong</td>
</tr>
<tr>
<td><strong>DEMO2VEC</strong> [12]</td>
<td>Manual affordance keypoints, action labels</td>
<td>Strong</td>
</tr>
</tbody>
</table>

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.
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 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 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.

Generalization to Novel Objects. Can interaction heatspots...
determine 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 $O$ into familiar and unfa-
miliar objects $O = O_f \cup O_u$; 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 $O_f$.
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 in-
teracted 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 out-
performs 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—

![Figure 5: Interaction hotspots on EPIC videos of unseen ob-
ject classes. Our model anticipates interaction regions for inactive
objects at rest (first column), before the interaction happens. Criti-
cally, the object categories shown in this figure were not seen dur-
ing 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.

which was never trained on some objects (e.g., cupboard, squash)—is able to anticipate characteristic spatial locations of interactions before the interaction occurs.

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 ob-
ject space (top branch) capture typical appearance-based vi-
sual 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 char-
acteristic way (top right); knives, spatulas, tongs, that are
typically held at their handles (bottom right). Our model
learns representations indicative of functional similarity be-
tween 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 experi-
ments show that these hotspot maps explain object afford-
dances better than other existing weakly supervised mod-
els and can generalize to anticipate affordances of unseen
objects. Furthermore, the representation learned by our
model goes beyond appearance similarity to encode func-
tional 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

[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, 11, 12, 14
[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


