EGO-TOPO: Environment Affordances from Egocentric Video

Tushar Nagarajan[1,2*] Yanghao Li[2] Kristen Grauman[1,2] Christoph Feichtenhofer[2]

[1]The University of Texas at Austin [2]Facebook AI Research (FAIR)

Abstract

First-person video naturally brings the use of a physical environment to the forefront, since it shows the camera wearer interacting fluidly in a space based on his intentions. However, current methods largely separate the observed actions from the persistent space itself. We introduce a model for environment affordances that is learned directly from egocentric video. The main idea is to gain a human-centric model of a physical space (such as a kitchen) that captures (1) the primary spatial zones of interaction and (2) the likely activities they support. Our approach decomposes a space into a topological map derived from first-person activity, organizing an ego-video into a series of visits to the different zones. Further, we show how to link zones across multiple related environments (e.g., from videos of multiple kitchens) to obtain a consolidated representation of environment functionality. On EPIC-Kitchens and EGTEA+, we demonstrate our approach for learning scene affordances and anticipating future actions in long-form video.

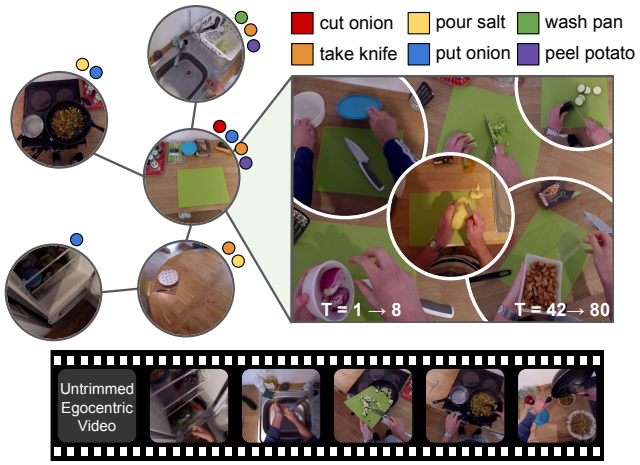

Figure 1: Main idea. Given an egocentric video, we build a topological map of the environment that reveals *activity-centric* zones and the sequence in which they are visited. These maps capture the close tie between a physical space and how it is used by people, which we use to infer affordances of spaces (denoted here with color-coded dots) and anticipate future actions in long-form video.

1. Introduction

"The affordances of the environment are what it offers the animal, what it provides or furnishes... It implies the complementarity of the animal and the environment."—James J. Gibson, 1979

In traditional third-person images and video, we see a moment in time captured intentionally by a photographer who paused to actively record the scene. As a result, scene understanding is largely about answering the *who/where/what* questions of recognition: what objects are present? is it an indoor/outdoor scene? where is the person and what are they doing? [55, 52, 72, 42, 73, 34, 69, 18].

In contrast, in video captured from a first-person "egocentric" point of view, we see the environment through the eyes of a person passively wearing a camera. The surroundings are tightly linked to the camera-wearer's ongoing interactions with the environment. As a result, scene understanding in egocentric video also entails *how* questions: how can one use this space, now and in the future? what areas are most conducive to a given activity?

Despite this link between activities and environments, existing first-person video understanding models typically ignore that the underlying environment is a persistent physical space. They instead treat the video as fixed-sized chunks of frames to be fed to neural networks [46, 5, 14, 65, 48, 41]. Meanwhile, methods that do model the environment via dense geometric reconstructions [63, 20, 57] suffer from SLAM failures—common in quickly moving head-mounted video—and do not discriminate between those 3D structures that are relevant to human actions and those that are not (*e.g.*, a cutting board on the counter versus a random patch of floor). We contend that neither the "pure video" nor the "pure 3D" perspective adequately captures the scene as an action-affording space.

Our goal is to build a model for an environment that captures how people use it. We introduce an approach called EGO-TOPO that converts egocentric video into a topological map consisting of activity "zones" and their rough spatial proximity. Taking cues from Gibson's vision above, each zone is a region of the environment that *affords a coherent set of interactions*, as opposed to a uniformly shaped

region in 3D space. See Fig. 1.

Specifically, from egocentric video of people actively using a space, we link frames across time based on (1) the physical spaces they share and (2) the functions afforded by the zone, regardless of the actual physical location. For example, for the former criterion, a dishwasher loaded at the start of the video is linked to the same dishwasher when unloaded, and to the dishwasher on another day. For the latter, a trash can in one kitchen could link to the garbage disposal in another: though visually distinct, both locations allow for the same action—discarding food. See Fig. 3.

In this way, we re-organize egocentric video into "visits" to known zones, rather than a series of unconnected clips. We show how doing so allows us to reason about first-person behavior (*e.g.*, what are the most likely actions a person will do in the future?) and the environment itself (*e.g.*, what are the possible object interactions that are likely in a particular zone, even if not observed there yet?).

Our EGO-TOPO approach offers advantages over the existing models discussed above. Unlike the "pure video" approach, it provides a concise, spatially structured representation of the past. Unlike the "pure 3D" approach, our map is defined organically by people's use of the space.

We demonstrate our model on two key tasks: inferring likely object interactions in a novel view and anticipating the actions needed to complete a long-term activity in first-person video. These tasks illustrate how a vision system that can successfully reason about scenes' functionality would contribute to applications in augmented reality (AR) and robotics. For example, an AR system that knows where actions are possible in the environment could interactively guide a person through a tutorial; a mobile robot able to learn from video how people use a zone would be primed to act without extensive exploration.

On two challenging egocentric datasets, EPIC and EGTEA+, we show the value of modeling the environment explicitly for egocentric video understanding tasks, leading to more robust scene affordance models, and improving over state-of-the-art long range action anticipation models.

2. Related Work

Egocentric video Whereas the camera is a bystander in traditional third-person vision, in first-person or egocentric vision, the camera is worn by a person interacting with the surroundings firsthand. This special viewpoint offers an array of interesting challenges, such as detecting gaze [40, 29], monitoring human-object interactions [4, 6, 51], creating daily life activity summaries [44, 39, 70, 43], or inferring the camera wearer's identity or body pose [28, 33]. The field is growing quickly in recent years, thanks in part to new ego-video benchmarks [5, 41, 54, 62].

Recent work to recognize or anticipate actions in egocentric video adopts state-of-the-art video models from third-person video, like two-stream networks [41, 46], 3DConv models [5, 53, 48], or recurrent networks [15, 16, 61, 65]. In contrast, our model grounds first-person activity in a persistent topological encoding of the environment. Methods that leverage SLAM together with egocentric video [20, 57, 63] for activity forecasting also allow spatial grounding, though in a metric manner and with the challenges discussed above, which we illustrate in our experiments.

Structured video representations Recent work explores ways to enrich video representations with more structure. Graph-based methods encode relationships between detected objects: nodes are objects or actors, and edges specify their spatio-temporal layout or semantic relationships (*e.g.*, is-holding) [67, 3, 45, 71]. Architectures for composite activity learn to encode atomic action "primitives" that are aggregated over the full time extent of the video [17, 30, 31], memory-based models record a recurrent network's state [53], and 3D convnets augmented with long-term feature banks provide temporal context [68]. Unlike any of the above, our approach encodes video in a human-centric manner according to how people use a space. In our graphs, nodes are spatial zones and connectivity depends on a person's visitation over time.

Mapping and people's locations Traditional maps use simultaneous localization and mapping (SLAM) to obtain dense metric measurements, viewing a space in strictly geometric terms. Instead, recent work in embodied visual navigation explores learning-based maps that leverage both visual patterns as well as geometry, with the advantage of extrapolating to novel environments (*e.g.*, [23, 22, 59, 26, 9]). Our approach shares this motivation. However, unlike any of the above, our approach analyzes egocentric video, as opposed to controlling a robotic agent. Furthermore, whereas existing maps are derived from a robot's exploration, our maps are derived from *human behavior*.

Work in ubiquitous computing tracks people to see where they spend time in an environment [36, 2], and "personal locations" manually specified by the camera wearer (*e.g.*, my office) can be recognized using supervised learning [12]. In contrast, our approach automatically discovers zones of activity from ego-video, and it links action-related zones across multiple environments.

Affordances Affordances are often focused on objects, where the goal is to anticipate how an object could be used—for example learning to model object manipulation [1, 4] or how people would grasp an object [37, 51, 10, 6]. People's body pose can even improve object recognition [7, 19]. The affordances of scenes are less studied. Prior work explores how a third-person view of a scene suggests likely 3D body poses that would occur there [60, 66, 21] and vice versa [11]. More closely re-

lated to our work, Action Maps [56] estimate missing activity labels for regular grid cells in an environment, using matrix completion with object and scene similarities as side information. In contrast, our work considers affordances not strongly tied to a single object's appearance, and we introduce a graph-based video encoding derived from our topological maps that benefits action anticipation.

3. EGO-TOPO Approach

We aim to organize egocentric video into a map of activity "zones"—regions that afford a coherent set of interactions—and ground the video as a series of visits to these zones. This representation offers a middle ground between the "pure video" and "pure 3D" approaches discussed above, which either ignore the underlying environment by treating video as fixed-sized chunks of frames, or sacrifice important semantics of human behavior by densely reconstructing the whole environment. Instead, our model reasons jointly about the environment and the agent: which parts of the environment are most relevant for human action, what interactions does each zone afford, and how actions at these zones accomplish a goal.

Our approach is best suited to long term activities in egocentric video where zones are repeatedly visited and used in multiple ways over time. This definition applies broadly to common household and workplace environments (*e.g.*, office, kitchen, retail store, grocery). In this work, we study kitchen environments using two public ego-video datasets (EPIC [5] and EGTEA+ [41]), since cooking activities entail frequent human-object interactions and repeated use of multiple zones. Our approach is not intended for third-person video, short video clips, or video where the environment is constantly changing (*e.g.*, driving down a street).

Our approach works as follows. First, we train a zone localization network to discover commonly visited spaces from egocentric video (Sec. 3.1). Then, given a novel video, we use the network to assign video clips to zones and create a topological map (graph) for the environment. We further link zones based on their function across video instances to create consolidated maps (Sec. 3.2). Finally, we train models that leverage the resulting graphs to uncover environment affordances (Sec. 3.3) and anticipate future actions in long videos (Sec. 3.4).

3.1. Discovering Activity-Centric Zones

We leverage egocentric video of human activity to discover important "zones" for action. At a glance, one might attempt to discover spatial zones based on visual clustering or geometric partitions. However, clustering visual features (*e.g.*, from a pretrained CNN) is insufficient since manipulated objects often feature prominently in ego-video, making the features sensitive to the set of objects present. For example, a sink with a cutting-board being washed vs. the same sink at a different time filled with plates would cluster into different zones. On the other hand, SLAM localization is often unreliable due to quick motions characteristic of egocentric video.[1] Further, SLAM reconstructs all parts of the environment indiscriminately, without regard for their ties to human action or lack thereof, *e.g.*, giving the same capacity to a kitchen sink area as it gives to a random wall.

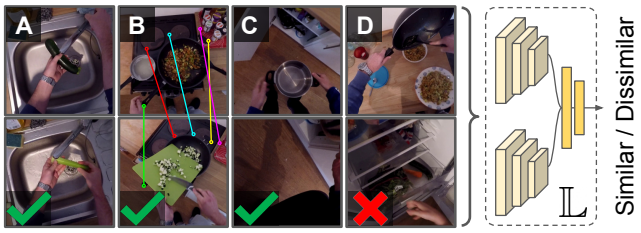

Figure 2: Localization network. Our similarity criterion goes beyond simple visual similarity (A), allowing our network to recognize the stove-top area (despite dissimilar features of prominent objects) with a consistent homography (B), or the seemingly unrelated views at the cupboard that are temporally adjacent (C), while distinguishing between dissimilar views sampled far in time (D).

To address these issues, we propose a zone discovery procedure that links views based on both their visual content and their visitation by the camera wearer. The basis for this procedure is a localization network that estimates the similarity of a pair of video frames, designed as follows.

We sample pairs of frames from videos that are segmented into a series of action clips. Two training frames are similar if (1) they are near in time (separated by fewer than 15 frames) or from the same action clip, or (2) there are at least 10 inlier keypoints consistent with their estimated homography. The former allows us to capture the spatial coherence revealed by the person's behavior and his/her tendency to dwell by action-informative zones, while the latter allows us to capture repeated backgrounds despite significant foreground object changes. Dissimilar frames are temporally distant views with low visual feature similarity, or incidental views in which no actions occur. See Fig. 2. We use SuperPoint [8] keypoint descriptors to estimate homographies, and euclidean distance between pretrained ResNet-152 [25] features for visual similarity.

The sampled pairs are used to train $\mathbb{L}$, a Siamese network with a ResNet-18 [25] backbone, followed by a 5 layer multi-layer perceptron (MLP), using cross entropy to predict whether the pair of views is similar or dissimilar. Once trained, the network can predict the probability $\mathbb{L}(f_t, f'_t)$ that two frames f_t, f'_t in an egocentric video belong to the same zone.

Our localization network draws inspiration from the retrieval network employed in [59] to build maps for embodied agent navigation, and more generally prior work lever-

[1] For example, on the EPIC Kitchens dataset, only 44% of frames can be accurately registered with a state-of-the-art SLAM algorithm [50].

Algorithm 1 Topological affordance graph creation.

Input: A sequence of frames $(f_1, ...f_T)$ of a video
Input: Trained localization network $\mathbb{L}$ (Sec. 3.1)
Input: Node similarity threshold σ and margin m

Create a graph $G = (N, E)$ with node $n_1 = \{(f_1 \rightarrow f_1)\}$
for $t \leftarrow 2$ to T **do**
  $s^* \leftarrow \max_{n \in N} s_f(f_t, n)$ — Equation 2
  if $s^* > \sigma$ **then**
    Merge f_t with node $n^* = \arg\max_{n \in N} s_f(f_t, n)$
    If f_t is a consecutive frame in n^*: Extend last visit v
    Else: Make new visit v with f_t
  else if $s^* < \sigma - m$ **then**
    Create new node, add visit with f_t, and add to G
  end if
  Add edge from last node to current node
end for

Output: EGO-TOPO topological affordance graph G per video

aging temporal coherence to self-supervise image similarity [24, 49, 32]. However, whereas the network in [59] is learned from view sequences generated by a randomly navigating agent, ours learns from ego-video taken by a human acting purposefully in an environment rich with object manipulation. In short, nearness in [59] is strictly about physical reachability, whereas nearness in our model is about human interaction in the environment.

3.2. Creating the Topological Affordance Graph

With a trained localization network, we process the stream of frames in a new untrimmed, unlabeled egocentric video to build a topological map of its environment. For a video $\mathcal{V}$ with T frames $(f_1, ..., f_T)$, we create a graph $G = (N, E)$ with nodes N and edges E. Each node of the graph is a zone and records a collection of "visits"—clips from the egocentric video at that location. For example, a cutting board counter visited at $t = 1$ and $t = 42$, for 7 and 38 frames each, will be represented by a node $n \in N$ with visits $\{v_1 = (f_1 \rightarrow f_8), v_2 = (f_{42} \rightarrow f_{80})\}$. See Fig. 1.

We initialize the graph with a single node n_1 corresponding to a visit with just the first frame. For each subsequent frame f_t, we compute the average *frame-level* similarity score s_f for the frame compared to each of the nodes $n \in N$ using the localization network from Sec. 3.1:

$$s_f(f_t, n) = \frac{1}{|n|} \sum_{v \in n} \mathbb{L}(f_t, f_v) \quad (1)$$

$$s^* = \max_{n \in N} s_f(f_t, n), \quad (2)$$

where f_v is the center frame selected from each visit v in node n. If the network is confident that the frame is similar to one of the nodes, it is merged with the highest scoring node n^* corresponding to s^*. Alternately, if the network is confident that this is a new location (very low s^*), a new node is created for that location, and an edge is created from the previously visited node. The frame is ignored if the network is uncertain about the frame. Algorithm 1 summarizes the construction algorithm. Further implementation details and values σ, m can be found in Supp.

Figure 3: Cross-map linking. Our linking strategy aligns multiple kitchens (P01, P13 etc.) by their common spaces (*e.g.*, drawers, sinks in rows 1-2) and visually distinct, but functionally similar spaces (*e.g.*, dish racks, crockery cabinets in row 3).

When all frames are processed, we are left with a graph of the environment per video where nodes correspond to zones where actions take place (and a list of visits to them) and the edges capture weak spatial connectivity between zones based on how people traverse them.

Importantly, beyond per-video maps, our approach also creates cross-video and cross-environment maps that link spaces by their function. We show how to link zones across 1) multiple episodes in the same environment and 2) multiple environments with shared functionality. To do this, for each node n_i we use a pretrained action/object classifier to compute $(\mathbf{a}_i, \mathbf{o}_i)$, the distribution of actions and active objects[2] that occur in all visits to that node. We then compute a *node-level* functional similarity score:

$$s_n(n_i, n_j) = -\frac{1}{2}\left(KL(\mathbf{a}_i||\mathbf{a}_j) + KL(\mathbf{o}_i||\mathbf{o}_j)\right), \quad (3)$$

where KL is the KL-Divergence. We score pairs of nodes across all kitchens, and perform hierarchical agglomerative clustering to link nodes with functional similarity. Details about the clustering algorithm are in Supp.

Linking nodes in this way offers several benefits. First, not all parts of the kitchen are visited in every episode (video). We link zones across different episodes in the *same* kitchen to create a *combined* map of that kitchen that accounts for the persistent physical space underlying multiple video encounters. Second, we link zones *across* kitchens to create a *consolidated* kitchen map, which reveals how different kitchens relate to each other. For example, a gas stove in one kitchen could link to a hotplate in another, despite being visually dissimilar (see Fig. 3). Being able to draw such parallels is valuable when planning to act in a new unseen environment, as we will demonstrate below.

[2] An active object is an object involved in an interaction.

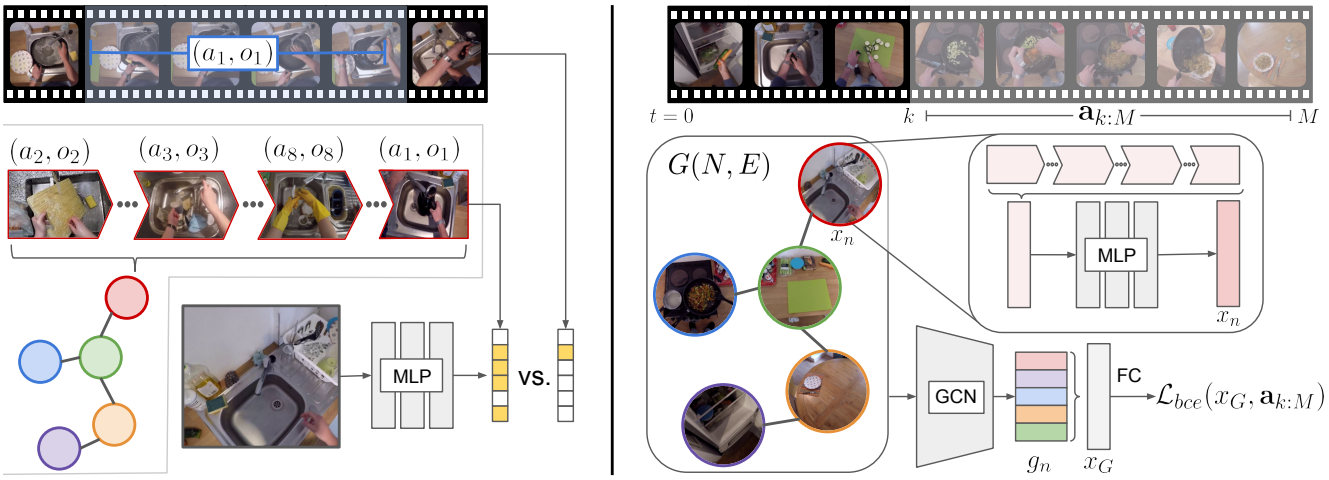

Figure 4: Our methods for environment affordance learning (L) and long horizon action anticipation (R). Left panel: Our EGO-TOPO graph allows multiple affordance labels to be associated with visits to zones, compared to single action labels in annotated video clips. Note that these visits may come from different videos of the same/different kitchen—which provides a more robust view of affordances (cf. Sec. 3.3). **Right panel**: We use our topological graph to aggregate features for each zone and consolidate information across zones via graph convolutional operations, to create a concise video representation for long term video anticipation (cf. Sec. 3.4).

3.3. Inferring Environment Affordances

Next, we leverage the proposed topological graph to predict a zone's affordances—*all likely* interactions possible at that zone. Learning scene affordances is especially important when an agent must use a previously unseen environment to perform a task. Humans seamlessly do this, *e.g.*, cooking a meal in a friend's house; we are interested in AR systems and robots that learn to do so by watching humans.

We know that egocentric video of people performing daily activities reveals how different parts of the space are used. Indeed, the actions observed per zone partially reveal its affordances. However, since each clip of an ego-video shows a zone being used only for a single interaction, it falls short of capturing *all likely* interactions at that location.

To overcome this limitation, our key insight is that linking zones within/across environments allows us to extrapolate labels for *unseen interactions* at *seen zones*, resulting in a more complete picture of affordances. In other words, having seen an interaction (a_i, o_i) at a zone n_j allows us to augment training for the affordance of (a_i, o_i) at zone n_k, if zones n_j and n_k are functionally linked. See Fig. 4 (Left).

To this end, we treat the affordance learning problem as a multi-label classification task that maps image features x_i to an A-dimensional binary indicator vector $\mathbf{y_i} \in \{0, 1\}^A$, where A is the number of possible interactions. We generate training data for this task using the topological affordance graphs $G(N, E)$ defined in Sec. 3.2.

Specifically, we calculate *node-level* affordance labels $\mathbf{y}_n$ for each node $n \in N$:

$$\mathbf{y}_n(k) = 1 \quad \text{for} \quad k \in \bigcup_{v \in n} \mathcal{A}(v) \tag{4}$$

where $\mathcal{A}(v)$ is the set of all interactions that occur during visit v. Then, from each visit to a node n, we sample a frame, generate its frame features x, and use $\mathbf{y}_n$ as the affordance multi-label target. We use a 2-layer MLP for the affordance classifier, followed by a linear classifier and a sigmoid function. The network is trained using binary cross entropy loss.

At test time, given an image x in an environment, this classifier directly predicts its affordance probabilities. See Fig. 4 (Left). Critically, linking frames into zones and linking zones between environments allows us to share labels across instances in a manner that that benefits affordance learning, better than models that link data purely based on geometric or visual nearness (cf. Sec. 4.1). Our EGO-TOPO graph derives its connectivity naturally by observing human use of related environments.

3.4. Anticipating Future Actions in Long Video

Next, we leverage our topological affordance graphs for long horizon anticipation. In the anticipation task, we see a fraction of a long video (*e.g.*, the first 25%), and from that we must predict what actions will be done in the future. Compared to affordance learning, which benefits from how zones are functionally related to enhance static image understanding, long range action anticipation is a video understanding task that leverages how objects are distributed among zones and how these zones are laid out to anticipate human behavior.

Recent action anticipation [13, 75, 15, 5, 53, 16, 61] predicts the immediate next action (*e.g.* in the next 1 second) rather than all future actions, for which an encoding of recent video information is sufficient. For long range antici-

pation, models need to first understand how much progress has been made on the composite activity so far, and then anticipate what actions need to be done in the future to complete it. For this, a structured representation of all past activity and affordances is essential.

Existing long range video understanding methods [30, 31, 68] build complex models over past clip features to aggregate information from the past, but do not model the environment explicitly, which we hypothesize is important for anticipating actions in long video. Our graphs provide a concise representation of observed activity, grounding frames in the spatial environment. We leverage this grounding to learn trends in interaction sequences.

Given an untrimmed video $\mathcal{V}$ with M interaction clips each involving an action $\{a_1, ..., a_M\}$ with some object, we see the first k clips[3] and predict the future action labels as a D-dimensional binary vector $\mathbf{a_{k:M}}$, where D is the number of action classes and $a^d_{k:M} = 1$ for $d \in \{a_{k+1}, ..., a_M\}$.

We generate the corresponding topological graph $G(N, E)$ built up to k clips, and extract features x_n for each node using a 2-layer MLP, over averaged clip features sampled from visits to that node.

Actions at one node influence future activities in other nodes. To account for this, we enhance node features by integrating neighbor node information from the topological graph using a graph convolutional neural network (GCN) [35]

$$g_n = ReLU\left(\sum_{n' \in \mathcal{N}_n} W^T x_{n'} + b\right), \tag{5}$$

where $\mathcal{N}_n$ are the neighbors of node n, and W, b are learnable parameters of the GCN.

The updated GCN representation g_n for each individual node is enriched with global scene context from neighboring nodes, allowing patterns in actions across locations to be learned. For example, vegetables that are *taken out* of the fridge in the past are likely to be *washed* in the sink later. The GCN node features are then averaged to derive a representation of the video $x_G = \frac{1}{|N|}\sum_{n \in N} g_n$. This is then fed to a linear classifier followed by a sigmoid to predict future action probabilities, trained using binary cross entropy loss, $\mathcal{L}_{bce}(x_G, \mathbf{a_{k:M}})$.

At test time, given an untrimmed, unlabeled video showing the onset of a long composite activity, our model can predict the actions that will likely occur in the future to complete it. Fig. 4 (Right) illustrates the task. Further details are in Supp. As we will see in results, grounding ego-video clips in the real environment—rather than treat them as an arbitrary set of frames—provides a stronger video representation for anticipation.

[3]Experiments sweep over values of k to test seeing more/less video.

4. Experiments

We evaluate the proposed topological graphs for scene affordance learning and action anticipation in long videos.

Datasets. We use two egocentric video datasets:

- **EGTEA Gaze+** [41] contains videos of 32 subjects following 7 recipes in a single kitchen. Each video captures a complete dish being prepared (*e.g.*, potato salad, pizza), with clips annotated for interactions (*e.g.*, open drawer, cut tomato), spanning 53 objects and 19 actions.
- **EPIC-Kitchens** [5] contains videos of daily kitchen activities, and is not limited to a single recipe. It is annotated for interactions spanning 352 objects and 125 actions. Compared to EGTEA+, EPIC is larger, unscripted, and collected across multiple kitchens.

The kitchen environment is an ideal setting for our experiments, and has been the subject of several recent egocentric datasets [5, 41, 38, 64, 58, 74]. Repeated interaction with different parts of the kitchen during complex, multi-step cooking activities is a rich domain for learning affordance and anticipation models.

4.1. EGO-TOPO for Environment Affordances

In this section, we evaluate how linking actions in zones and across environments can benefit affordances.

Baselines. We compare the following methods:

- **CLIPACTION** uses clip-level action labels to learn to recognize those afforded actions which it has seen at a given location during training.
- **ACTIONMAPS** [56] estimates affordances of locations via matrix completion with side-information. It assumes that nearby locations with similar appearance/objects have similar affordances. See Supp. for details.
- **SLAM** trains an action affordance classifier with the same architecture as ours, and treats all frames associated with the same grid cell on the ground plane as positives for actions observed at any time in that grid cell. (x, y) locations are obtained from monocular SLAM [50], and the cell size is based on the typical scale of an interaction area following prior work [20]. It shares our insight to link actions in the same location, but is limited to a uniformly defined location grid and cannot link different environments.
- **KMEANS** clusters action clips using their visual features alone. We select as many clusters as there are nodes in our consolidated graph to ensure fair comparison.
- **OURS** We show the three variants from Sec. 3.2 which use maps built from a single video (OURS-S), multiple videos of the same kitchen (OURS-M), and a functionally linked, consolidated map across kitchens (OURS-C).

	EPIC			EGTEA+		
mAP →	ALL	FREQ	RARE	ALL	FREQ	RARE
CLIPACTION	26.8	49.7	16.1	46.3	58.4	33.1
ACTIONMAPS [56]	21.0	40.8	13.4	43.6	52.9	31.3
SLAM	26.6	48.6	17.6	41.8	49.5	31.8
KMEANS	26.7	50.1	17.4	49.3	**61.2**	35.9
OURS-S	28.6	52.2	19.0	48.9	61.0	35.3
OURS-M	28.7	53.3	18.9	**51.6**	**61.2**	**37.8**
OURS-C	**29.4**	**54.5**	**19.7**	–	–	–

Table 1: Environment affordance prediction. Our method outperforms all other methods. Note that videos in EGTEA+ are from the same kitchen, and do not allow cross-kitchen linking. Values are averaged over 5 runs.

Note that all methods use the clip-level annotated data, in addition to data from linking actions/spaces. They see the same video frames during training, only they are organized and presented with labels according to the method.

We crowd-source annotations for afforded interactions. Each instance is a view from the environment, paired with *all* likely interactions at that location regardless of whether the view shows it (*e.g.*, turn-on stove, take/put pan etc. at a stove). We collect 1020 instances spanning $A = 75$ interactions on EGTEA+ and 1155 instances over $A = 120$ on EPIC (see Supp. for details). All methods are evaluated on this test set. We report mean average precision (mAP) over all afforded interactions, and separately for the rare and frequent ones ($<$10 and $>$100 training instances, respectively).

Table 1 summarizes the results. By capturing the persistent environment in our discovered zones, and linking them across environments, our method outperforms all other methods on the affordance prediction task. All models perform better on EGTEA+, which has fewer interaction classes, contains only one kitchen, and has at least 30 training examples per afforded action (compared to EPIC where 10% of the actions have a single annotated clip).

SLAM and ACTIONMAPS [56] rely on monocular SLAM, which introduces certain limitations. See Fig. 5 (Left). A single grid cell in the SLAM map reliably registers only small windows of smooth motion, often capturing only single action clips at each location. In addition, inherent scale ambiguities and uniformly shaped cells can result in incoherent activities placed in the same cell. Note that this limitation stands even if SLAM were perfect. Together, these factors hurt performance on both datasets, more severely affecting EGTEA+ due to the scarcity of SLAM data (only 6% accurately registered). Noisy localizations also affect the kernel computed by ACTIONMAPS, which accounts for physical nearness as well as similarities in object/scene features. In contrast, a zone in our topological affordance graph corresponds to a coherent set of clips at different times, offering a more reliable and diverse set of actions to link, as seen in Fig. 5 (Right).

Clustering using purely visual features in KMEANS helps consolidate information in EGTEA+ where all videos

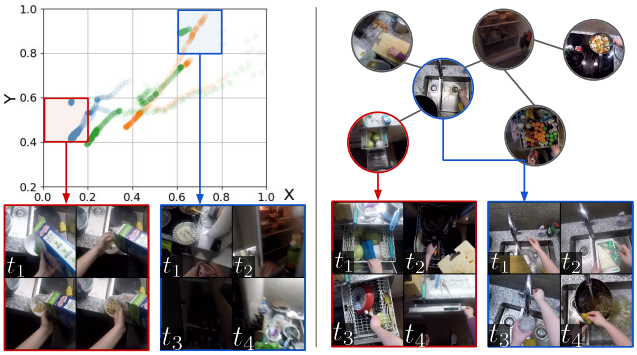

Figure 5: SLAM grid vs graph nodes. The boxes below show frames from video that are linked to grid cells in the SLAM map (Left) and nodes in our topological map (Right). See text.

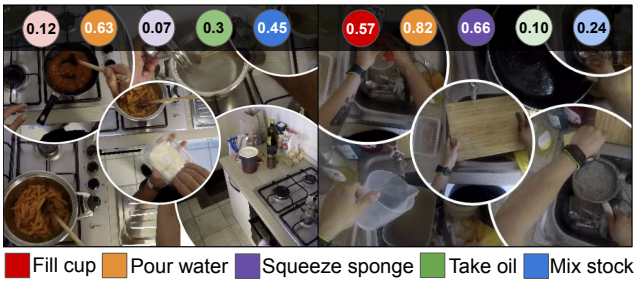

Figure 6: Predicted affordance scores for graph nodes. Our affordance model applied to node visits reveal zone affordances.

are in the same kitchen, but hurts performance where visual features are insufficient to capture coherent zones.

Linking actions to discovered zones in our topological graph results in consistent improvements on both datasets. Moreover, aligning spaces based on function in the consolidated graph (OURS-C) provides the largest improvement, especially for rare classes that may only be seen tied to a single location.

Fig. 3 and Fig. 5 show the diverse actions captured in each node of our graph. Multiple actions at different times and from different kitchens are linked to the same zone, thus overcoming the sparsity in demonstrations and translating to a strong training signal for our scene affordance model. Fig. 6 shows example affordance predictions.

4.2. EGO-TOPO for Long Term Action Anticipation

Next we evaluate how the structure of our topological graph yields better video features for long term anticipation.

Baselines. We compare against simple strategies to aggregate clip features, as well as state-of-the art methods for long range temporal aggregation [17, 30, 31].

- **TRAINDIST** simply outputs the distribution of actions performed in all training videos, to test if a few dominant actions are repeatedly done, regardless of the video.
- **I3D** uniformly samples 64 clip features and averages them to generate a video feature.

	EPIC			EGTEA+		
mAP →	ALL	FREQ	RARE	ALL	FREQ	RARE
TRAINDIST	16.5	39.1	5.7	59.1	68.2	35.2
I3D	32.7	53.3	23.0	72.1	79.3	53.3
RNN	32.6	52.3	23.3	70.4	76.6	54.3
ACTIONVLAD [17]	29.8	53.5	18.6	73.3	79.0	58.6
VIDEOGRAPH [31]	22.5	49.4	14.0	67.7	77.1	47.2
TIMECEPTION [30]	35.6	55.9	26.1	**74.1**	79.7	**59.7**
OURS W/O GCN	34.6	55.3	24.9	72.5	79.5	54.2
OURS	**38.0**	**56.9**	**29.2**	73.5	**80.7**	54.7

Table 2: Long term anticipation results. Our method outperforms all others on EPIC, and is best for many-shot classes on the simpler EGTEA+. Values are averaged over 5 runs.

- **RNN** and **ACTIONVLAD** [17] model temporal dynamics in video using LSTM [27] layers and non-uniform pooling strategies, respectively.
- **TIMECEPTION** [30] and **VIDEOGRAPH** [31] build complex temporal models using either multi-scale temporal convolutions or attention mechanisms over learned latent concepts from clip features over large time scales.

While all the compared methods model temporal information, none explicitly model the persistent environment in video as we propose.

For evaluation, we use the first $K\%$ of each untrimmed video as input, and predict all actions in the remaining video. We sweep values of $K = [25\%, 50\%, 75\%]$ representing different anticipation horizons. We report mAP over all action classes, and in low-shot (rare) and many-shot (freq) settings.

Table 2 shows the results averaged over all K's, and Fig. 7 plots results vs. K. Our model outperforms all other methods on EPIC, improving over the next strongest baseline by 2.4% mAP on all 125 action classes. On EGTEA+, our model matches the performance of models with complicated temporal aggregation schemes, and achieves the highest results for many-shot classes.

EGTEA+ has a less diverse action vocabulary with a fixed set of recipes. TRAINDIST, which simply outputs a fixed distribution of actions for every video, performs relatively well (59% mAP) compared to its counterpart on EPIC (only 16% mAP), highlighting that there is a core set of repeatedly performed actions in the dataset.

Among the methods that employ complex temporal aggregation schemes, TIMECEPTION improves over I3D on both datasets, though our method outperforms it on the larger EPIC dataset. Simple aggregation of node level information (OURS W/O GCN) still consistently outperforms most baselines. However, including the graph convolution operations is essential to outperform more complex models, which shows the benefit of encoding the physical layout and interactions between zones in our topological map.

Fig. 7 breaks down performance by anticipation horizon K. On EPIC, our model is uniformly better across

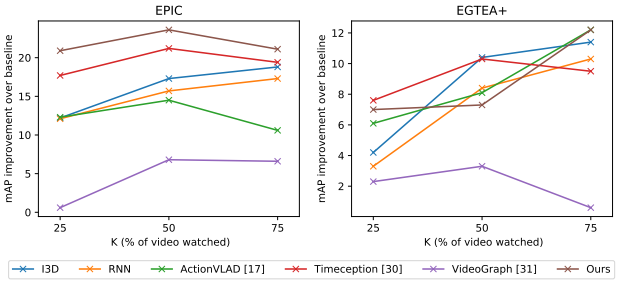

Figure 7: Anticipation performance over varying prediction horizons. $K\%$ of the video is observed, then the actions in the remaining $100-K\%$ must be anticipated. Our model outperforms all methods for all anticipation horizons on EPIC, and has higher relative improvements when predicting further into the future.

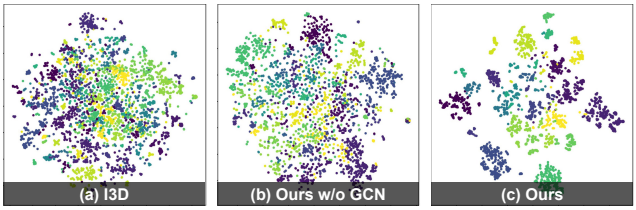

Figure 8: t-SNE visualization on EPIC. (a) Clip-level features from I3D; Node features for OURS (b) without and (c) with GCN. Colors correspond to different kitchens.

all prediction horizons, and it excels at predicting actions further into the future. This highlights the benefit of our environment-aware video representation. On EGTEA+, our model outperforms all other models except ACTIONVLAD on short range settings, but performs slightly worse at K=50%. On the other hand, ACTIONVLAD falls short of all other methods on the more challenging EPIC data.

Fig. 8 shows a t-SNE[47] visualization of the learned feature spaces. Though trained for anticipation, features from our method discriminate between different kitchens in EPIC, without explicit kitchen labels, to encode useful environment-aware information.

5. Conclusion

We proposed a method to produce a topological affordance graph from egocentric video of human activity, highlighting commonly used zones that afford coherent actions across multiple kitchen environments. Our experiments on scene affordance learning and long range anticipation demonstrate its viability as an enhanced representation of the environment gained from egocentric video. Future work can leverage the environment affordances to guide users in unfamiliar spaces with AR or allow robots to explore a new space through the lens of how it is likely used.

References

[1] J.-B. Alayrac, J. Sivic, I. Laptev, and S. Lacoste-Julien. Joint discovery of object states and manipulation actions. *ICCV*,

2017. 2

[2] D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with gps. In *Proceedings Sixth International Symposium on Wearable Computers*, pages 101–108. IEEE, 2002. 2

[3] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018. 2

[4] M. Cai, K. Kitani, and Y. Sato. Understanding hand-object manipulation with grasp types and object attributes. In *RSS*, 2016. 2

[5] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2, 3, 5, 6

[6] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. 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

[7] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012. 2

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 3

[9] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 538–547, 2019. 2

[10] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 2

[11] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. *IJCV*, 2014. 2

[12] A. Furnari, S. Battiato, and G. M. Farinella. Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. *Journal of Visual Communication and Image Representation*, 52:1–12, 2018. 2

[13] A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 5

[14] A. Furnari and G. M. Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[15] A. Furnari and G. M. Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. *arXiv preprint arXiv:1905.09035*, 2019. 2, 5

[16] J. Gao, Z. Yang, and R. Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2, 5

[17] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–980, 2017. 2, 7, 8

[18] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 1

[19] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR*, 2011. 2

[20] J. Guan, Y. Yuan, K. M. Kitani, and N. Rhinehart. Generative hybrid representations for activity forecasting with no-regret learning. *arXiv preprint arXiv:1904.06250*, 2019. 1, 2, 6

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

[22] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2017. 2

[23] S. Gupta, D. Fouhey, S. Levine, and J. Malik. Unifying map and landmark based representations for visual navigation. *arXiv preprint arXiv:1712.08125*, 2017. 2

[24] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 4

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[26] J. F. Henriques and A. Vedaldi. Mapnet: An allocentric spatial memory for mapping environments. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018. 2

[27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 8

[28] Y. Hoshen and S. Peleg. An egocentric look at video photographer identity. In *CVPR*, 2016. 2

[29] Y. Huang, M. Cai, Z. Li, and Y. Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, 2018. 2

[30] N. Hussein, E. Gavves, and A. W. Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 2, 6, 7, 8

[31] N. Hussein, E. Gavves, and A. W. Smeulders. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*, 2019. 2, 6, 7, 8

[32] D. Jayaraman and K. Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *CVPR*, 2016. 4

[33] H. Jiang and K. Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3509. IEEE, 2017. 2

[34] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene

graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1

[35] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 6

[36] K. Koile, K. Tollmar, D. Demirdjian, H. Shrobe, and T. Darrell. Activity zones for context-aware computing. In *International Conference on Ubiquitous Computing*, pages 90–106. Springer, 2003. 2

[37] H. S. Koppula and A. Saxena. Physically grounded spatio-temporal object affordances. In *ECCV*, 2014. 2

[38] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 6

[39] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal on Computer Vision*, 114(1):38–55, August 2015. 2

[40] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013. 2

[41] Y. Li, M. Liu, and J. M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018. 1, 2, 3, 6

[42] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[43] C. Lu, R. Liao, and J. Jia. Personal object discovery in first-person videos. *IEEE Trans. on Image Processing*, 24(12), Dec 2015. 2

[44] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[45] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018. 2

[46] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 1, 2

[47] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[48] A. Miech, I. Laptev, J. Sivic, H. Wang, L. Torresani, and D. Tran. Leveraging the present to anticipate the future in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[49] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *ICML*, 2009. 4

[50] M. J. M. M. Mur-Artal, Raúl and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 3, 6

[51] T. Nagarajan, C. Feichtenhofer, and K. Grauman. Grounded human-object interaction hotspots from video. *arXiv preprint arXiv:1812.04558*, 2018. 2

[52] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *2011 International Conference on Computer Vision*, pages 1307–1314. IEEE, 2011. 1

[53] F. Pirri, L. Mauro, E. Alati, V. Ntouskos, M. Izadpanahkakhk, and E. Omrani. Anticipation and next action forecasting in video: an end-to-end model with memory. *arXiv preprint arXiv:1901.03728*, 2019. 2, 5

[54] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2847–2854. IEEE, 2012. 2

[55] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009. 1

[56] N. Rhinehart and K. M. Kitani. Learning action maps of large environments via first-person vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–588, 2016. 3, 6, 7

[57] N. Rhinehart and K. M. Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 1, 2

[58] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201. IEEE, 2012. 6

[59] N. Savinov, A. Dosovitskiy, and V. Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018. 2, 3, 4

[60] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner. Scenegrok: Inferring action maps in 3d environments. *TOG*, 2014. 2

[61] Y. Shi, B. Fernando, and R. Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 2, 5

[62] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 2

[63] H. Soo Park, J.-J. Hwang, Y. Niu, and J. Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016. 1, 2

[64] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738. ACM, 2013. 6

[65] S. Sudhakaran, S. Escalera, and O. Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019. 1, 2

[66] X. Wang, R. Girdhar, and A. Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2596–2605, 2017. 2

[67] X. Wang and A. Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018. 2

[68] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 2, 6

[69] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 1

[70] R. Yonetani, K. M. Kitani, and Y. Sato. Visual motif discovery via first-person vision. In *ECCV*, 2016. 2

[71] Y. Zhang, P. Tokmakov, M. Hebert, and C. Schmid. A structured model for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9975–9984, 2019. 2

[72] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1

[73] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 1

[74] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 6

[75] Y. Zhou and T. L. Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 5