Multi-label affordance mapping from egocentric vision

Lorenzo Mur-Labadia Jose J. Guerrero Ruben Martinez-Cantin
I3A - Universidad de Zaragoza
lmur, jguerrer, rmcantin@unizar.es

Abstract

Accurate affordance detection and segmentation with pixel precision is an important piece in many complex systems based on interactions, such as robots and assitive devices. We present a new approach to affordance perception which enables accurate multi-label segmentation. Our approach can be used to automatically extract grounded affordances from first person videos of interactions using a 3D map of the environment providing pixel level precision for the affordance location. We use this method to build the largest and most complete dataset on affordances based on the EPIC-Kitchen dataset, EPIC-Aff, which provides interaction-grounded, multi-label, metric and spatial affordance annotations. Then, we propose a new approach to affordance segmentation based on multi-label detection which enables multiple affordances to co-exists in the same space, for example if they are associated with the same object. We present several strategies of multi-label detection using several segmentation architectures. The experimental results highlight the importance of the multi-label detection. Finally, we show how our metric representation can be exploited for build a map of interaction hotspots in spatial action-centric zones and use that representation to perform a task-oriented navigation.

1. Introduction

When humans repeatedly interact in a close environment, we associate a set of affordable actions with a certain distribution of objects in a physical space. For example, we associate a pan on a stove with cooking, but the same pan on the sink with washing. A joined spatial-semantic understanding contains powerful insights to understand human behaviour. This requires a close combination of perception, mapping and navigation algorithms; with potential applications in augmented reality systems [60, 61], but also guiding a robot [19, 36] or assistive devices [68]. In the last years, the ability of deep learning models to extract high-level representations has improved the perception of autonomous agents, while egocentric vision offers a powerful viewpoint for modelling human-object interaction understanding. Recent advances include anticipating future actions [23, 26, 1], model the hands-object manipulation [22, 77, 27, 16], detect the change in an object state [2], identify interaction hotspots [20, 54] or create topological maps [55]. Despite the fast movements of a headset camera, egocentric perception has also contributed to the mapping and planning phases: localising the agent in a known 3D map [43], performing visual navigation [56, 62] or building *third-person* (allocentric) maps [7, 47].

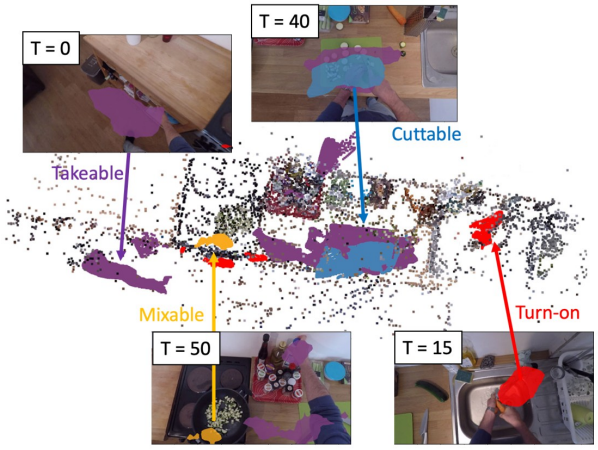

Figure 1. From a sequence of egocentric observations, our agent creates a spatial-metric multi-label representation of the affordances, enabling a task-oriented navigation.

Gibson's perception theory presents affordances as the potential actions that the environment offers to the agent based on its motor capabilities [25]. For example, the person can afford *taking* a glass, but the affordances of a soup in a pan can be *mixing, emptying, scooping* and *pouring* simultaneously. This multiplicity models better complex dynamic environments and opens the door to multi-agent collaboration with task synchronization. Although some authors have focused on more complex affordance models [50, 75], affordance perception is typically defined as a classification problem. Some authors have focused instead on grounded affordances [20], which provide a more flexible setup and are truly associated with motor capabilities, show-

ing improvements in action anticipation [46]. However, most learning approaches in affordance perception consider the problem ungrounded to the agent interaction with the object, requiring previous annotations of each affordance occurrence [53, 58, 48, 19, 8, 57]. While ungrounded methods have the advantage of providing pixel-wise precision, which we call metric understanding of the scene, many grounded approaches rely on full image classification losing any metric meaning. In this paper, we propose a grounded approach with pixel-wise precision, which enables detailed metric understanding while maintaining the flexibility of grounded methods and that can be used as prior information for more complex affordance models [75]. Close to our proposal is the work of Nagarajan et al. [54], which presented a grounded approach for extracting interaction hotspots by directly observing videos. Similar to other previous works, the hotspots are modelled as a single available affordance. Instead, we propose to consider the multiplicity of affordances for a single object or spatial zone through multi-label pixel-wise predictions.

We build a pipeline to automatically collect multi-label pixel-wise annotations from real-world interactions using a temporal, spatial and semantic representation of the environment. We use this method with the EPIC Kitchens videos [15] to build a dataset of grounded affordances (EPIC-Aff), which to the author's knowledge, constitute the largest dataset in affordance segmentation up to date. We then adapt several segmentation architectures to the multi-label paradigm to extract more diverse information from the scene based on the assumption that the same object may have multiple affordances available. Using a mapping approach we extract the multi-label affordance segmentation to build a map that spatially links activity-centric zones as shows Figure 1, allowing a metric representation of the environmental affordances and goal-directed navigation tasks. Finally, we perform a quantitative evaluation of the extension from common architectures to the multi-label paradigm and we show mapping and planning applications of our approach, that can be used for assistive devices or in robotic scenarios.

2. Related works

2.1. Learning Visual Affordances

Ungrounded approaches [53, 58, 48, 19, 8, 57] for affordance perception are fully supervised by manually labelled masks. Due to their similarity with semantic segmentation or object detection tasks, these works use common architectures such as an encoder-decoder [57], proposal-based detectors [19, 58, 48, 8] or Bayesian instance segmentation [52]. Concerning grounded works, Fang et al. [20] extracted a latent representation from demonstration videos or Luo et al. [46] transferred the learning from exocentric images to the egocentric perspective using only the semantic label as supervision. Nagarajan et al. [54] extracted the interaction hotspots by deriving the gradient-weighted attention maps obtained at training an action classifier on videos. Then Ego-Topo [55] built a topological graph of the scene to perform affordance classification from egocentric videos, grouping each node visually and temporally coherent frames with similar object and action distributions. This allows them to discover activity-centric zones based on their visual content and represent semantically the traversed paths with the edges of the graph.

2.2. Multi-label perception

In the multi-label segmentation task, we assign two or more categories to a single pixel. A particular case is an amodal segmentation where the relevance of occluded parts depends on the depth order [78, 49]. The most common applications of multi-label segmentation are biomedical works, where there are multiple overlapped non-exclusive levels of tissues. Existing architectures extend a U-Net with minor changes such as a dynamic segmentation head [18], shuffle-attention mechanisms in the skip-connections [39], a combination of appearance and pose features [5] and split-attention modules [38]. The closest approach to multi-label segmentation is multi-label image recognition, where the label imbalance between positives and negatives in each binary classifier and the extraction of features from multiple objects make this task more complex [45]. Class distribution aware losses [73] such as the asymmetric loss [66], the focal loss [42] or the Multi-Label Softmax loss [24] correct the over-suppression of negative samples. On the other hand, Graph Neural Networks such as [74] deal with the feature extraction from multiple objects by creating a dynamic graph for each image that leverages the content-aware category representations. Finally, the transformer architecture extracts multiple attention maps in the different regions of interest [13, 37], guiding the multi-label classification [45] or ranking the class of the pixels considering only the categories selected by the classifier [30].

3. Grounded Affordance Labelling

We extract *automatic*, *interaction-grounded*, *multi-label pixel-wise* and *spatial* affordance annotations from a sequence of real-world images in complex and cluttered environments, as shows Figure 2. Our multi-label segmentation approach learns *all* the potential options and does not reduce the perception to a single action. For example, a potato on a chopping board offers *cutting, putting, peeling, removing* and *taking* simultaneously. Current affordance segmentation works [19, 53, 58, 46, 52] assume a single-label affordance per object and lose a valuable amount of information. Although other affordance models allow for multiple predictions, these works ignore the segmentation of the inter-

Dataset	Year	IG	Pix	ML	CP	#Obj.	#Aff.	#Imgs.
UMD [53]	2017	X	✓	X	X	17	7	30,000
IIT-Aff [58]	2017	X	✓	✓	X	10	9	8,835
ADE-Aff [14]	2018	X	✓	✓	X	150	3	10,000
OPRA [20]	2018	✓	✓	X	X	-	7	20,774
Grounded I.H [54]	2018	✓	✓	X	X	31	20	1,800*
Ego-Topo [55]	2020	✓	X	✓	X	304	75-120	1,020-1,115*
PAD v2 [76]	2021	X	✓	X	X	72	31	30,000
AGD20k [46]	2022	X	✓	X	X	50	36	23,816
EPIC-Aff	2023	✓	✓	✓	✓	304	20-43	38,876

Table 1. **Visual affordance datasets statistics**. I.G: Interaction Grounded. Pix: pixel-wise annotations. ML: multi-label. CP: camera poses #Obj: Number of objects. #Aff: Number of affordances. #Imgs: total number of images. * The affordance labels are only for evaluation, the model is trained supervised only by action labels provided by [15]

action hotspot in the image and lose the pixel-level accuracy of the segmentation models. For example, topological maps extract multiple affordances from an image [55], or action anticipation models predict a probability distribution of the different possibilities [23, 26]. Our methodology gets the best of two worlds producing multi-label metric masks, resulting in a full distribution of affordances. It enables a deeper understanding of the manipulation task such as the grasping points of the tool [3] or the evolution of the manipulation process over time[44]. Similar to previous unsupervised or weakly supervised methods [50], we extract affordance labels from weak VISOR and EPIC Kitchens annotations grounded on actual interactions.

We join the affordances with their 3D spatial location by extracting the camera poses. The spatial approach to affordance perception is not new for the community. Rhinehart et al. [64] associate the functionality of regions with specific spatial locations, showing that that defining an affordance based solely on semantics is insufficient due to the significant influence of the physical context. For example, a frying pan is only *cookable* when it is on the hob or a plate is *washable* when the agent is next to the dishwasher. However, their method results in smooth 2D maps, which can be problematic for the fine-grained affordances in 3D space in our EPIC-Aff dataset. Instead, our method is able to scale up to large environments while maintaining the detail by using neural networks. Other previous works [28, 65] also use SLAM for action prediction but with addressing different problems. In those works, the action is set on the human, while the image provides context; while in our case the action/affordance is set on the environment and the user provides context. In our work, we use COLMAP to extract the relative pose between sparse frames with a filter of the dynamic objects, registering up to 93 % of the frames compared with the 44 % of the frames registered by ORB-SLAM [51] on EPIC Kitchens [55]. Recently, EPIC Fields [72] registered the camera pose of the dense videos in EPIC Kitchens using neural rendering techniques.

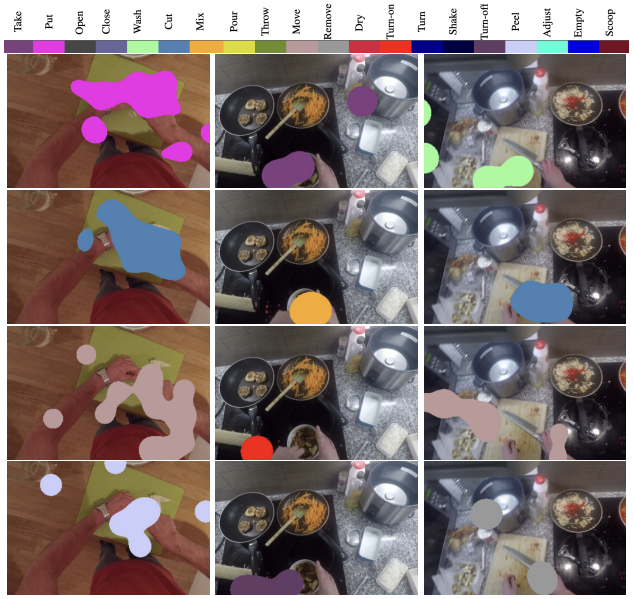

Figure 2. Ground truth examples. For visualization purposes, we show a single label of the affordable action on its location, although these are overlapped for the same sample. The food in the bowl affords *taking* or *mixing*, while the cutting board on the left affords *putting*, *cutting*, *moving* and *peeling*.

3.1. Affordance datasets

Following our motivation, we conduct a study along the visual affordance datasets shown in Table 1. The ungrounded datasets are subjected to the annotator's consideration and required to draw pixel-wise semantic mask to each object part [58, 53, 14, 76, 46] or additional sensors [35, 9], decreasing the object variability and limiting the scalability due to the annotation costs. The UMD dataset contains a semantic affordance map for objects in isolated conditions and with low variability, which prevents generalisation [53]. The IIT-Aff dataset [58] provides the most comprehensive annotations designed for use in robotics, including multiple objects in a single image. The ADE-Aff dataset [14], built on top of ADE20K scenes, examines the social acceptability of actions about context but is limited to only three affordance classes. The AGD20k [46] dataset includes the largest number of categories and actions by transferring from an exocentric to an egocentric viewpoint perspective. On the other hand, grounded works learn from observing interactions on the EPIC-Kitchens sequences [15], internet demonstration videos [20] or with gaze point with eye-tracking devices [21]. The annotations provided are only used for evaluation since they do not require strong supervision. However, these approaches ignore the pixel-wise precision [55] or the multi-label modality of our approach[54].

Based on the mentioned limitations, our novel dataset EPIC-Aff provides multi-label pixel-wise affordance annotations with the camera pose. It contributes to a diverse and

comprehensive affordance database with the largest number of images up to date. This better captures the complexity, dynamics, multiplicity and variability of real-world environments, such as preparing a recipe in a kitchen. Finally, as our labels are automatically extracted, we enable the application of our method to other egocentric datasets.

3.2. EPIC-Aff dataset

We detail the procedure shown in Figure 3 for our grounded affordance labelling. EPIC-Aff[1] is composed of 38,876 images with up to 43 different affordable actions $\mathcal{K}$. We choose the EPIC-Kitchens as the base dataset because of its sequential and repetitive nature, which allows us to extract the 3D geometry, and because the kitchen is a scenario with multi-step and structured activities very rich in semantics. We cover all the object categories present in the EPIC-100 annotations, which constitute a wide, large and diverse knowledge base.

From a sequence of video, we join the EPIC-100 narrations [15] and the VISOR Kitchen annotations [17] to obtain a sparse sequence of frames $\mathcal{S}_{\mathcal{M}} = (f_1, ..., f_N)$ with the localization of the interactions on the image, as shows Figure 4. The EPIC-100 labels [15] contains narrations formed by an action verb $\mathcal{V}$ with an associated object $\mathcal{O}$, i.e: "add steak", for more than 100 hours of video. VISOR Kitchens [17] interpolates from sparse annotations to generate semantic masks $\mathcal{M}$ and bounding boxes $\mathcal{B}$ on the active objects. We set the centre of the interaction $x_i = \{u_i, v_i\}$ in the middle of the intersection between the hand $\mathcal{B}_h$ and the interacted object $\mathcal{B}_O$ bounding-boxes given by the narration $\mathcal{V} + \mathcal{O}$.

Then, we apply COLMAP, a Structure-from-Motion (SfM) algorithm [69] to obtain the camera poses T_w^c and a point cloud of the environment $\{X_p\}$. In the EPIC-Kitchens [15], each kitchen is composed of multiple videos, thus, we join all the sparse frames with interactions $\mathcal{S}_K = \{\mathcal{S}_1 \dots \mathcal{S}_M\}$ to relate frames from different videos in a common 3D reference. Furthermore, we use the VISOR semantic masks to avoid including dynamic objects in the point cloud.

Next, we employ a robust depth estimator based on a neural network $d_{NN} = f_d(\cdot)$ [63] to predict the depth of interaction points $d_{NN}(x_i)$. Because the neural network computes the depth up-to-scale, we compute a scale correction factor per image to fit the network scale to the SfM scale: $scale = median(d^{SfM}(X_p)/median(d^{NN}(X_p))$ [34], where $d^{SfM}(X_p)$ is the depth of all the points $\{X_p\}$ visible from the current image and $d^{NN}(X_p)$ is the depth of the same points given by the network estimator.

Using the predicted depth and the scale projection, we can project the interaction point x_i in 3D space X_i and use the camera pose to project into the global coordinates $X_i^w = T_w^c \cdot X_i^c$. At this point, as shown in Figure 5 we obtain in a common reference a history of all the interactions that occured in the kitchen $\mathcal{I}_k = \{X_1^w, X_2^w, \dots, X_k^w\}$, cross-generalizing for the different sequences. This constitutes our knowledge base that follows our hypothesis that the distribution of affordances is spatially linked to pre-determined physical spaces (*i.e* you only wash in the dishwasher), not only to the semantic context of a topological graph [55].

Then, once we store all the past interactions $\mathcal{I}_k$ with their $\mathcal{V}_k$ and $\mathcal{O}_k$ labels, we reproject them back to the new camera reference system $X_i^c = T_c^w \cdot X_i^w$. Instead of considering all the object semantic masks as the affordance region, we centre a Gaussian distribution over each affordance reprojected point X_i^c and build an additive heatmap. Then, the affordance masks $\mathcal{M}_i^{aff}$ are defined as the regions where the heatmap is greater than 0.25. This is grounded in how humans interact with objects [54] and allows us to consider the different affordability of the object parts (a knife is only *graspable* with the handle). In order to generate the affordance labels $\mathcal{A} = \{(\mathcal{V}_1, \mathcal{O}_1, \mathcal{M}_1^{aff}), ..., (\mathcal{V}_j, \mathcal{O}_j)\}$, we select only those verbs whose associated object $\mathcal{O}_i$ was present in the VISOR annotations $\{\mathcal{M}, \mathcal{B}\}$. With this procedure, we are grounding our dataset in the past interactions in that environment and associating multiple affordances to a single object. We show qualitative samples of the EPIC-Aff in Figure 2.

We provide two different versions of the dataset: the easy EPIC-Aff and the complex-EPIC Aff, with 20 and 43 affordance classes respectively. There is a challenging class imbalance, as shown the Figure 6 with a significant frequency gap between the most common class (*open*, with a 16.0 %) and the less represented (*dry*, with a 0.3 %). In Fig. 7 we show the pixel ratio, which reflects that semantically similar or opposite actions are associated with the same space (*i.e, turn-on, turn-off, adjust* or *cut* and *peel*) and the importance of the multi-level approach. This shows that activity-centric zones are physical spaces where there occur multiple common activities, both synonyms or antonyms. For example, in the hob controls region we likely find (*i.e, turn-on, turn-off, adjust)*, while on the dishwasher zone, we will encounter *wash* and *dry*.

4. Multi-label segmentation and mapping

In this section, we explain our inference procedure. First, we describe the modifications needed to obtain a multi-label segmentation model. We then show how our approach can be applied to mapping and planning tasks.

4.1. Multi-label segmentation

In this section we describe how we transform classical semantic segmentation models to a multi-label version. While there exists lots of single-label segmentation

[1] https://github.com/lmur98/epic_kitchens_affordances

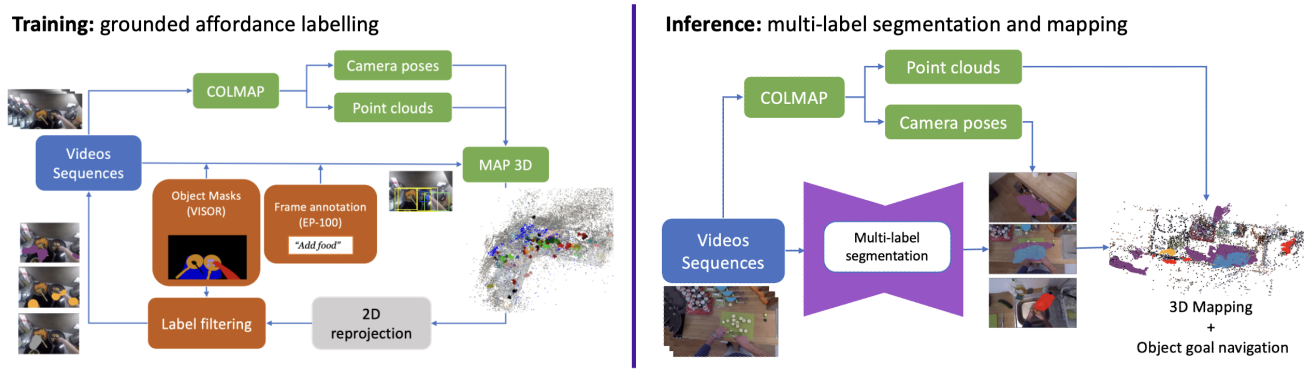

Figure 3. Left: Pipeline with the automatic extraction of the pixel-wise labels on the EPIC-Aff dataset. We combine the EPIC-100 narration with the VISOR masks annotations to extract the interaction point. Then, using the camera pose extracted from COLMAP, we project all the interactions in a common 3D global reference. Finally, we reproject all the past interactions to each frame, and filter the affordance annotation by the objects present at the image. Right: the multi-label masks predictions from our model are leveraged to a 3D map

Figure 4. Using the masks provided by VISOR Kitchens, we define the intersection between the object and the hand bounding boxes as the center of the interaction. We show in yellow the bounding box of the non-interacting objects, in green the bounding box of the hands and in blue the bounding box of the interacting object.

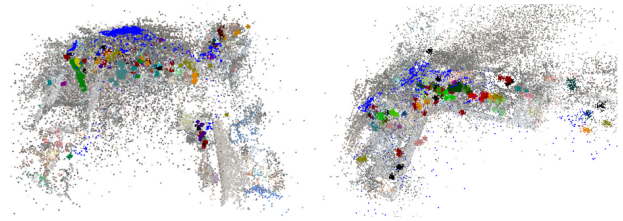

Figure 5. Historical with all the past interaction frames in that environment, where the blue dots represent the camera poses of the sparse frames from all the sequences

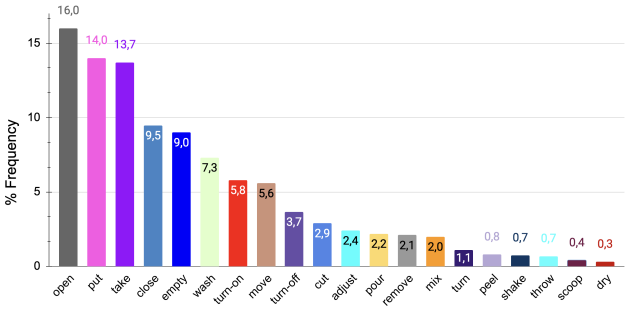

Figure 6. Distribution of the 20 classes in the easy-EPIC Aff dataset, showing a significant class imbalance.

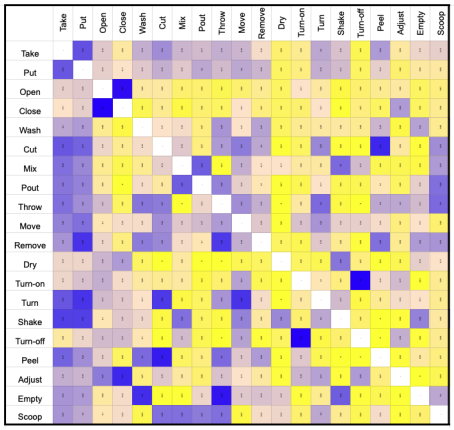

Figure 7. Pixel ratio of the 20 classes in the easy-EPIC Aff dataset, where *blue* represents high correlation between the two classes and *yellow* that they do not used to concur in the same pixel.

[11, 40, 4, 31, 30, 10] and multi-label image classification works [73, 24, 12, 45], the multi-label segmentation is a more unexplored task restricted to small domains like biomedical images [39, 5].

Given an input image $\mathbf{X}$, the multi-label segmentation goal is to predict a group of categories for each pixel. Therefore, we assume that each pixel could represent multiple affordances (*takeable*, *cuttable*, *washable*, ..., etc.) or not belong to any category. For a total number of $\mathcal{K}$ classes we define the label y for each pixel of the image as $y = [y_1, ..., y_k]$, where $y_k = 1$ if the pixel contains the $\mathcal{K}$-category label, otherwise $y_k = 0$. In order to predict multi-label segmentation masks, we have evaluated two different approaches. First, we use a standard multiclass segmentation networks and evaluate three different heuristics to select multiple labels per pixel. Then, we modify the segmentation networks to output multiple binary classifiers which enable multiple labels to be active.

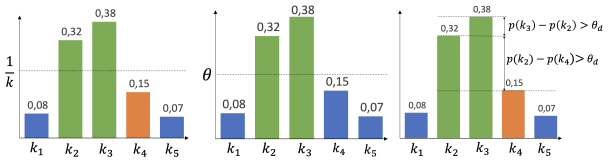

Figure 8. Heuristics to select multiple labels from a probability vector

For the multiclass scenario, we assume that the network output is a categorical distribution for all the classes and use the standard supervision loss, the cross-entropy. Then, we transform the probability vector $p = [p_1, ..., p_k]$, $\sum_{k=1}^{k} p_k = 1$ with three heuristics to choose the multiple winning-classes, as shows Figure 8. On the first method, we select the top-k classes with the largest probability value p_k. Note that we do not considers predictions with a $s_k < 1/k$, as it occurs to k_1, k_5 on Figure 8. The second alternative is max-θ, which consists in selecting all the possible classes whose p_k is greater than a threshold θ. Finally, the last heuristic is a dynamic θ_d threshold. We select the classes whose difference with the next class is larger than a θ_d.

On the multi-label scenario, the model outputs $\mathcal{K}$ independent Bernouilli distributions, generating binary probabilities $p = [p_1, ..., p_k]$, where we assume a detection if $p_k > 0.5$. Then, we substitute the cost function by a class-weighting Binary Cross Entropy (BCE) loss, obtaining $\mathcal{K}$ binary classifiers. One disadvantage of having independent binary classifiers is that the performance is more sensitive to the class imbalance in the dataset. To alleviate that, we use the Asymmetric (Asym) loss $\mathcal{L}_{asym}$ [66] shown in Equation 1. It combines the focal loss [71] with the margin loss [70] to reduce the contribution of easy negative samples and rejects mislabeled samples with a continuous gradient.

$$
\begin{aligned}
ASL_k &= \begin{cases} \log(p_k)(1-p_k)^{\gamma+}, & y_k = 1 \\ \log(1-p_k)(p_k)^{\gamma-}, & y_k = 0 \end{cases} \\
\mathcal{L}_{asym} &= \frac{1}{N} \sum^{N} \frac{w_k}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} ASL_k
\end{aligned}
\tag{1}
$$

For each training image, $\mathbf{X}$ with N total pixels, $\mathcal{L}_{asym}$ computes a different term depending on if the y_k binary label indicating that the class k is present or not in the pixel. We apply a weighting average w_c depending on the ratio between positive and negative samples for class k to avoid the class imbalance. Following the original paper [66], we set $\gamma^+ = 4$ and $\gamma^- = 1$.

4.2. Example applications

Given that our system provides metric information of the affordance location, has information of the camera poses and has multi-label affordance detection, we can apply it to common spatial tasks such as mapping and navigation.

Mapping of activity-centric zones We take the video sequence and sample unseen frames $\{f_1, ..., f_t\}$ during training. Following the same procedure as in the extraction of labels, we reproject the inferred semantic masks on the pixels i, j to its respective 3D location x, y, z using the camera intrinsic K_{int}, COLMAP pose R_w^c, t_w^c and the scaled depth $d_{i,j}$.

$$
\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d_{i,j} (R_w^c)^{-1} K_{int}^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} - t_w^c \tag{2}
$$

We accumulate in a global map the COLMAP key-points to represent the geometry and the segmented affordances regions. We do not perform any fusion on voxels or octrees, since our multi-label approach assumes that a zone can represent several predicted classes. Note that the map representation is common for all the sequences of the same environment, with the potential of linking zones across multiple episodes and learning from past interactions.

Task-oriented navigation Finally, we introduce a task-oriented navigation experiment to show the relevance of the map representation. We use the COLMAP key-points to build an occupancy grid with the available free space. Then, the agent is initialized in a random localization and asked to navigate to perform a certain action. Once it selects the location from the semantic-metric representation, the agent decides the path planning using a A* search with the Euclidean distance on the free space. We use the point cloud from COLMAP to create an occupancy grid.

5. Experiments

5.1. Models and metrics

In our experiments, we modify three popular semantic segmentation architectures [33, 67, 10] and compare them with a instance segmentation model [31] plus an interaction hotspots model [54].

- Grounded Interaction Hotspots (GIH) [54]: We use the weights on its EPIC-Kitchen trained version to extract predictions from our images. To reduce the gap, we crop our scenes to represent a single object and compare for the same number of affordable actions $\mathcal{K}$ in the easy-EPIC Aff dataset.

- Mask-RCNN [31]: we assume an overlapping in the bounding boxes between two different instances. We do not consider the amodal Mask-RCNN versions [59, 49] which treat differently visible and occlusion

	Take	Put	Open	Close	Wash	Cut	Mix	Pour	Throw	Move	Remove	Dry	Turn-on	Turn	Shake	Turn-off	Peel	Adjust	Empty	Scoop	mIoU
GIH [54]	22.1	21.9	13.8	10.8	16.3	25.8	21.2	23.7	14.0	16.9	17.2	12.8	10.3	20.5	16.6	10.8	26.1	9.5	13.9	25.8	17.5
Mask R-CNN [31]	**37.7**	**36.9**	**47.1**	**43.9**	**51.5**	41.4	**46.4**	38.1	43.6	**42.9**	38.4	13.1	**52.5**	43.7	30.8	**50.9**	35.3	47.1	33.6	26.0	40.1
U-Net dyn-θ[67]	0.1	0.7	5.4	11.9	22.4	17.1	22.2	17.3	11.3	15.9	21.0	4.5	14.8	21.1	16.3	18.4	12.9	20.6	0.5	9.6	13.2
(ours) U-Net BCE	22.3	22.5	30.9	24.0	30.2	23.7	21.1	17.7	17.1	23.4	18.5	8.7	27.3	22.4	13.2	23.8	16.2	22.2	19.0	13.1	20.9
(ours) U-Net Asym	14.3	13.7	13.8	14.7	21.3	17.9	18.3	18.7	32.5	15.7	15.6	16.6	15.2	18.9	22.2	19.5	19.5	24.3	5.7	15.7	17.7
FPN dyn-θ [33]	2.4	2.4	5.6	10.2	21.7	13.2	17.7	17.0	11.5	13.6	20.0	4.6	13.8	22.6	12.5	15.5	14.4	17.3	0.8	9.7	12.9
(ours) FPN BCE	25.7	25.9	33.3	26.7	33.2	22.4	21.8	15.4	18.5	23.9	21.4	7.4	3.8	20.8	13.3	28.4	13.6	24.5	23.5	11.6	22.2
(ours) FPN Asym	36.3	34.7	46.1	42.0	46.8	42.7	42.2	37.5	43.3	41.7	**39.6**	21.3	47.4	43.7	34.3	45.0	33.8	46.3	**38.0**	33.2	39.8
Deep-Lab v3 dyn-θ [10]	10.1	11.0	15.4	17.3	19.1	19.4	25.2	19.1	14.7	17.9	17.1	9.2	20.4	31.9	25.3	26.5	24.3	31.7	18.0	18.1	19.5
(ours) Deep-Lab v3 BCE	33.3	34.2	44.1	37.6	43.1	32.0	30.9	26.2	28.9	33.7	27.6	13.2	41.6	27.6	22.2	39.1	22.3	35.1	32.3	20.7	31.3
(ours) Deep-Lab v3 Asym	31.6	32.9	37.3	37.8	44.5	**43.9**	45.0	**41.8**	**53.4**	42.3	39.4	**33.1**	45.5	**52.2**	**44.0**	46.7	**43.5**	**51.1**	32.3	**46.6**	**42.3**

Table 2. Class-wise IoU scores on easy-EPIC Aff test set. All scores are in [%].

	KLD ↓	SIM ↑	AUC-J ↑	mIoU ↑	F1-Score ↑	mAP ↑	AP50 ↑
GIH [54]	2.381	0.116	0.511	17.5	29.4	14.2	15.5
Mask-RCNN [31]	1.365	0.150	0.841	40.1	56.5	**59.3**	**62.6**
U-Net [67] top-$\mathcal{K}$	2.532	0.341	0.830	9.5	17.4	22.0	30.5
U-Net [67] max-θ	2.532	0.341	0.830	13.2	23.6	22.0	30.5
U-Net [67] dyn-θ	2.532	0.341	0.830	13.2	23.7	22.0	30.5
(ours) U-Net + BCE	2.718	0.304	0.949	20.9	34.2	48.2	44.7
(ours) U-Net + Asym	0.783	0.665	0.857	17.7	29.9	15.6	32.3
FPN [33] top-$\mathcal{K}$	2.229	0.362	0.812	8.9	15.6	18.9	24.7
FPN [33] max-θ	2.229	0.362	0.812	12.4	21.8	18.9	24.7
FPN [33] dyn-θ	2.229	0.362	0.812	12.9	23.6	18.9	24.7
(ours) FPN + BCE	1.613	0.365	0.955	22.2	35.7	48.7	44.5
(ours) FPN + Asym	0.789	0.546	0.956	39.8	56.8	44.1	59.3
DeepLab-v3 [10] top-$\mathcal{K}$	4.947	0.192	0.911	18.9	31.9	35.0	40.9
DeepLab-v3 [10] max-θ	4.947	0.192	0.911	19.2	32.3	35.0	40.9
DeepLab-v3 [10] dyn-θ	4.947	0.192	0.911	19.5	32.7	35.0	40.9
(ours) DeepLab-v3 + BCE	1.276	0.179	0.964	31.3	47.2	58.6	56.2
(ours) DeepLab-v3 + Asym	**0.603**	**0.668**	**0.965**	**42.3**	**60.1**	43.6	58.5

Table 3. Affordance multi-label segmentation on easy-EPIC Aff test set (20 classes). Note that except the mIoU and the F1-Score, the rest of the metrics are common for the three versions of the multi-class segmentation models.

	KLD ↓	SIM ↑	AUC-J ↑	mIoU ↑	F1-Score ↑	mAP ↑	AP50 ↑
Mask-RCNN	2.287	0.211	0.756	17.1	27.3	**40.1**	**46.7**
(ours) U-Net Asym	1.104	0.320	0.657	12.9	24.8	11.2	17.9
(ours) FPN Asym	0.530	**0.673**	0.921	28.1	42.9	24.8	43.4
(ours) DeepLab-v3 Asym	**0.520**	0.670	**0.931**	**31.1**	**46.5**	27.4	43.9

Table 4. Affordance multi-label segmentation on complex-EPIC Aff test set (43 classes).

masks, since our affordance classes $\mathcal{K}$ are not ranked in order.

- Semantic segmentation architectures. We compare the performance of UNet [67], Feature Pyramid Networks (FPN) [33, 41] and DeepLab v-3 [10].

We train the segmentation models with an input resolution of 232 $\times$ 348 for 100 k iterations usign Adam as optimizer with weight decay of 10^{-4}, batch size of 8 and a initial learning rate of 10^{-4}, using a polynomial decay up to 10^{-6}. We apply random crop, color jitter, resize and flipping as data augmentation. In the same way, we train Mask-RCNN SGD and 10^{-2} as initial learning rate. We use a Resnet-50 backbone pre-trained on Imagenet for all the models in order to perform a fair comparative.

Following the evaluation of Nagarajan et al, [54], we report the Kullback-Leibler Divergence (KLD) [20], the Similarity metric (SIM) and the Area Under the Curve (AUC-J) [6, 32] which provide different metrics for the mismatch of

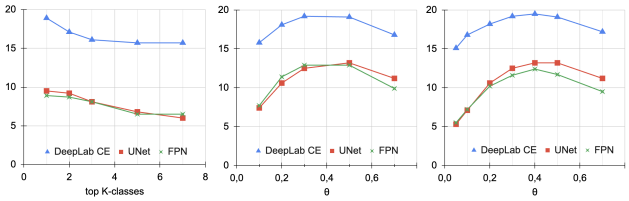

Figure 9. Evolution of the mIoU for different heuristics to select multiple winning classes from a multi-class probability vector. Left: top-$\mathcal{K}$. Center: max-θ. Right: dyn-θ.

the distribution of heatmaps or affordance regions considering the predictive probability. We also report metrics from segmentation literature, such as the mean Intersection over the Union (mIoU) and the F1-Score to measure the performance of the semantic segmentation, and the Average Precision (AP) AP-50 and mAP to report the performance of the detection metrics.

5.2. Quantitative results

We compare the performance of different popular architectures on the multi-label affordance segmentation task in Table 3 on the easy-EPIC Aff dataset. DeepLab-v3 trained with the Asymmetric loss obtains the best performance on the segmentation and saliency metrics (42.3 % mIoU 60.1 % F-1 score, 0.603 KLD, 0.668 SIM, 0.965 AUC-J). Since the backbone of the three semantic segmentation models is the same (Resnet-50), the different results are due to the configuration of the different decoders. In the dataset, since the labels represent interaction hotspots, they are not aligned with the borders of the objects and represent a more high-level zone. Thus, the atrous convolution of DeepLab enables to enlarge of the filter's field of view and better captures these regions. We show the per-class segmentation performance in terms of the IoU in Table 2. Comparing with the apparition frequency of the classes in the dataset shown in Figure 6, Mask R-CNN fails at the low-represented classes since is not trained with the Asymmetric loss. However, it is the best architecture on the detection metrics (59.3 % mAP, 62.6 % AP50). Compared with previous works, the pre-trained version of [54] achieves inter-

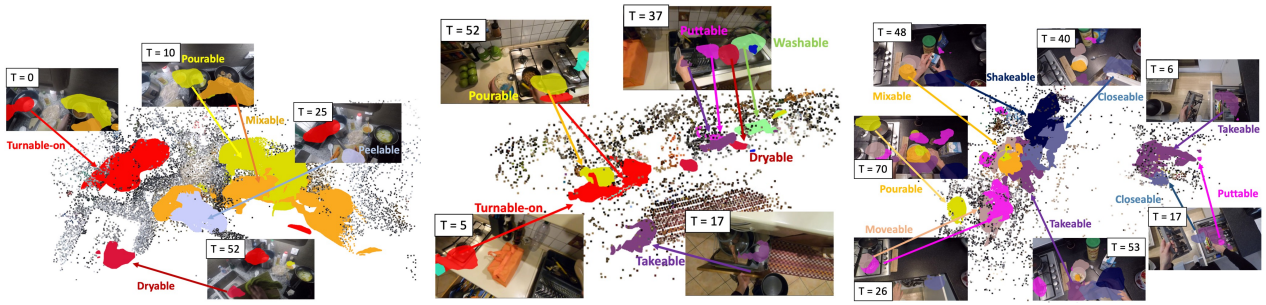

Figure 10. Spatial distribution of the detected multi-label affordances for multiple time-steps

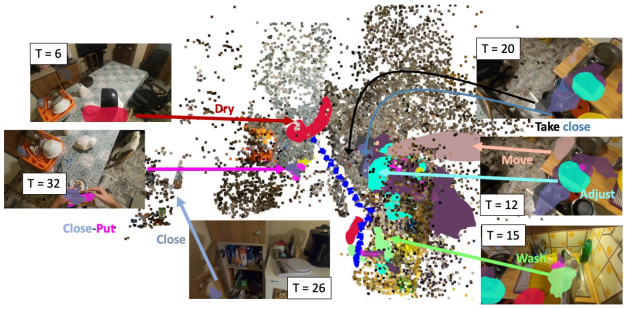

Figure 11. Goal oriented path-planning. In the example, at t=36 we indicate the user the trajectory from the sink to the place where it used to dry the crockery. The blue points represents the steps of the path planning

mediate results on the segmentation metrics (17.5 % mIoU, 29.4 % F-1 score) but low on the AP scores.

The results in Figure 9 show the impact of the hyper-parameters when adapting the multi-class models. The top-$\mathcal{K} = 1$ represents the classical multi-label case. The results show how its performance is far from the multi-label versions, supporting the need for specific architectural changes for this scenario. In Figure 9 left, when we increase the number of winning classes, the performance decreases by introducing too many false positives. The other two heuristics achieve better performance since they better reject these outliers. For example, the dyn-θ adapts dynamically to the probability distribution shape, obtaining higher mIoU and F-1 in the three cases (see Table 3). Finally, we appreciate similar results on the complex-EPIC Aff, shown in Table 4. In this case, the overall performance decreases due to the higher number of classes and its imbalance.

5.3. Mapping: metric distribution of affordances

We show on Figure 10 qualitative results of the multi-label interaction hotspots from affordances predicted by the DeepLab-v3 Asym model. Our perception model is consistent from different view-points. For example, the microwave of the left-map in Figure 10 is detected as *turn-on* both at $t = 0, 25, 52$. The qualitative results clearly motivate our multi-label approach: the milkshake on the right map affords *mixing*, *pouring* and *taking*, or the sink in the center map affords *drying* and *washing*. The metric conception of our approach is also relevant, since it reflects the interactions hotspots rather than highlight the complete object (for example *grasping* a pan only with the handle).

5.4. Task-oriented navigation

Finally, we use the spatial localization of the affordances to show a proof-of-concept "task-oriented" navigation. As we illustrate on Figure 11, we guide the agent according to the action possibilities that the environment offers to him. Therefore, we can ask our system to *perform* certain action, meaning to *go to where the object and affordance are available*. The A* indicates to the agent the shortest path from its current location to the position where it took the action in the past. For example, this could guide a visually impaired person with an assistant device [29].

6. Limitations

Our current approach presents several limitations. At the dataset extraction, we assume that the interaction occurs in the intersection between the object and the hand bounding-boxes, thus it depends on the bounding-box aligned to the actual object. This could be mitigated with a detection model for grasping points, but we wanted a simpler version for our prototype as a more convoluted approach might introduce further biases, difficult to detect. Also, the camera poses from COLMAP can be distorted by noisy-frames or dynamic objects non-suppressed by the mask. Furthermore, a real-time mapping system would require a SLAM system such as ORB-SLAM [51] which might reduce the accuracy of COLMAP. Similarly, our dataset is fully based on Kitchen sequences and it does not incorporate another environments introducing important dataset bias in the trained models. However, our automatic labeling pipeline could be easily used to extend the dataset in other scenarios.

7. Conclusions

We introduced a novel multi-label, metric and spatial-oriented perception of affordances. First, we present a

method for extracting grounded affordances labels based on egocentric interaction videos through a common metric representation of all the past interactions in a common reference. We use this pipeline to build the most complete affordance dataset based on the classic EPIC-Kitchen dataset. This constitutes EPIC-Aff, the largest semantic segmentation dataset of affordances grounded on the human interactions. We also motivate a method for grounded affordance detection with pixel precission using multi-label predictors, which enhances the perception and the representation of the environment. Furthermore, we show that the metric representation obtained can be used to build detailled affordance maps and to guide the user to perform task-oriented navigation tasks.

Acknowledgements

This work was supported by the Spanish Government (PID2021-125209OB-I00, TED2021-129410B-I00 and TED2021-131150B-I00) and the Aragon Government (DGA-T45_23R).

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5343–5352, 2018.

[2] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2127–2136, 2017.

[3] Paola Ardón, Èric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[5] Savinien Bonheur, Darko Štern, Christian Payer, Michael Pienn, Horst Olschewski, and Martin Urschler. Matwo-capsnet: a multi-label semantic segmentation capsules network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2019.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[7] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic mapnet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[8] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Are standard object segmentation models sufficient for learning affordance segmentation? *arXiv preprint arXiv:2107.02095*, 2021.

[9] Claudio Castellini, Tatiana Tommasi, Nicoletta Noceti, Francesca Odone, and Barbara Caputo. Using object affordances to improve object recognition. *IEEE transactions on autonomous mental development*, 3(3):207–215, 2011.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[11] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[12] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[13] Xing Cheng, Hezheng Lin, Xiangyu Wu, Dong Shen, Fan Yang, Honglin Liu, and Nian Shi. MLTR: multi-label classification with transformer. In *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*, pages 1–6. IEEE, 2022.

[14] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018.

[15] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

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

[17] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Ely Locke Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[18] Ruining Deng, Quan Liu, Can Cui, Zuhayr Asad, Yuankai Huo, et al. Single dynamic network for multi-label renal pathology image segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 304–314. PMLR, 2022.

[19] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018.

[20] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018.

[21] Alireza Fathi, Yin Li, James M Rehg, et al. Learning to recognize daily actions using gaze. *ECCV (1)*, 7572:314–327, 2012.

[22] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017.

[23] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019.

[24] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem. *arXiv preprint arXiv:1807.07247*, 2018.

[25] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[26] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021.

[27] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3293–3303, 2022.

[28] Jiaqi Guan, Ye Yuan, Kris M Kitani, and Nicholas Rhinehart. Generative hybrid representations for activity forecasting with no-regret learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 173–182, 2020.

[29] João Guerreiro, Daisuke Sato, Dragan Ahmetovic, Eshed Ohn-Bar, Kris M Kitani, and Chieko Asakawa. Virtual navigation for blind people: Transferring route knowledge to the real-world. *International Journal of Human-Computer Studies*, 135:102369, 2020.

[30] Haodi He, Yuhui Yuan, Xiangyu Yue, and Han Hu. Rankseg: Adaptive pixel classification with image category ranking for segmentation. In *European Conference on Computer Vision*, pages 682–700. Springer, 2022.

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[32] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[33] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019.

[34] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–713, 2018.

[35] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 831–847. Springer, 2014.

[36] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[37] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021.

[38] Minho Lee, JeeYoung Kim, Regina EY Kim, Hyun Gi Kim, Se Won Oh, Min Kyoung Lee, Sheng-Min Wang, Nak-Young Kim, Dong Woo Kang, ZunHyan Rieu, et al. Split-attention u-net: a fully convolutional network for robust multi-label segmentation from brain mri. *Brain Sciences*, 10(12):974, 2020.

[39] Michael Lempart, Martin P Nilsson, Jonas Scherman, Christian Jamtheim Gustafsson, Mikael Nilsson, Sara Alkner, Jens Engleson, Gabriel Adrian, Per Munck af Rosenschöld, and Lars E Olsson. Pelvic u-net: multi-label semantic segmentation of pelvic organs at risk for radiation therapy anal cancer patients using a deeply supervised shuffle attention convolutional neural network. *Radiation Oncology*, 17(1):1–15, 2022.

[40] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[41] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[43] Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *European Conference on Computer Vision*, pages 621–638. Springer, 2022.

[44] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.

[45] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[46] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2252–2261, 2022.

[47] Daniel Maturana, Po-Wei Chou, Masashi Uenoyama, and Sebastian Scherer. Real-time semantic mapping for autonomous off-road navigation. In *Field and Service Robotics: Results of the 11th International Conference*, pages 335–350. Springer, 2018.

[48] Chau Nguyen Duc Minh, Syed Zulqarnain Gilani, Syed Mohammed Shamsul Islam, and David Suter. Learning affordance segmentation: An investigative study. In *2020 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2020.

[49] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21023–21032, 2022.

[50] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.

[51] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[52] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose.J Guerrero. Bayesian deep learning for affordance segmentation in images. In *2023 International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[53] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[54] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019.

[55] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020.

[56] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egocentric scene context for human-centric environment understanding from video. *arXiv preprint arXiv:2207.11365*, 2022.

[57] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Detecting object affordances with convolutional neural networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2765–2770. IEEE, 2016.

[58] 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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017.

[59] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.

[60] Rodrigo Chacón Quesada and Yiannis Demiris. Holo-spok: Affordance-aware augmented reality control of legged manipulators. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 856–862. IEEE, 2022.

[61] Rodrigo Chacón Quesada and Yiannis Demiris. Proactive robot assistance: Affordance-aware augmented reality user interfaces. *IEEE Robotics & Automation Magazine*, 29(1):22–34, 2022.

[62] Santhosh Kumar Ramakrishnan, Tushar Nagarajan, Ziad Al-Halah, and Kristen Grauman. Environment predictive coding for visual navigation. In *International Conference on Learning Representations*, 2021.

[63] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[64] Nicholas Rhinehart and Kris 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.

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

[66] Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 82–91. IEEE, 2021.

[67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[68] Melani Sanchez-Garcia, Ruben Martinez-Cantin, and Jose J Guerrero. Semantic and structural image segmentation for prosthetic vision. *Plos one*, 15(1):e0227677, 2020.

[69] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[70] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International*

Conference on Machine Learning, pages 6105–6114. PMLR, 2019.

[71] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[72] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *arXiv preprint arXiv:2306.08731*, 2023.

[73] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

[74] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.

[75] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2155–2163, 2023.

[76] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500, 2022.

[77] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 127–145. Springer, 2022.

[78] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017.