2HandedAfforder: Learning Precise Actionable Bimanual Affordances from Human Videos

Marvin Heidinger[*1], **Snehal Jauhri**[*1], **Vignesh Prasad**[1], **Georgia Chalvatzaki**[1,2]
[*] indicates equal contribution
[1]Computer Science Department, Technische Universität Darmstadt, Germany
[2]Hessian.AI, Darmstadt, Germany
{marvin.heidinger, snehal.jauhri, vignesh.prasad}@tu-darmstadt.de,
georgia.chalvatzaki@tu-darmstadt.de

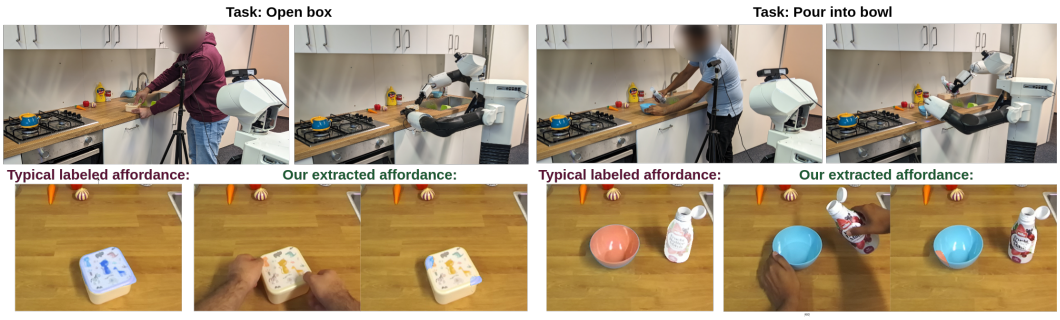

Figure 1: Examples of our affordance extraction in a real-world setting. The humans perform the tasks 'open box' and 'pour into bowl'. With our affordance extraction method, we obtain the precise, actionable object regions used, which leads to downstream success when the robot executes the skill.

Abstract: Robots need to perform tasks by interacting with and manipulating objects in their environment. Affordance prediction involves predicting the correct regions of objects to interact with to perform a task. Existing methods simplify this problem to naive object part segmentation. We propose a framework to extract and predict 'real' and 'actionable' affordances from videos of humans performing or demonstrating tasks. We propose the 2HANDS dataset of actionable affordances & show the performance of a baseline model to learn meaningful and actionable affordances.

Keywords: Affordance detection, Egocentric vision, Bimanual Robots

1 Introduction

It is a dream of many to deliver the 'household robot', i.e., a robot that can perform useful tasks in everyday, unstructured household environments. An essential requirement for such robots is the ability to interact with and manipulate objects in environments such as a kitchen, a living room, etc. This requires the robot to understand affordance regions of objects, i.e., the regions of objects that should be interacted with to perform a task. For example, to pour into a bowl, the robot should know that it should hold the bottle in a region in the middle of the bottle (Figure 1), i.e., a region that *affords* pouring. Predicting such affordance regions is challenging since it requires fine-grained understanding of object regions and their semantic relationship with the task. Moreover, the robot may also need to adapt affordance predictions to different humans' needs. For example, one end-user of the robot may prefer the robot to hold their bottle from a different region than another user.

Recent advances in large vision-language and multimodal models have shown impressive visual reasoning capabilities using self-supervised objectives [1, 2, 3]. However, there is still a big gap regarding their ability to detect object affordance regions in images [4] accurately. Moreover, most

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

existing state-of-the-art affordance detection methods [5, 6, 7, 8, 9] use labeled data [5, 6, 10, 11, 12] that lacks precision and is more akin to object part segmentation rather than *actionable* affordance region prediction. When humans interact with objects, they are much more *precise* and use specific object regions important for the task. An example is provided in Figure 1. For the task of opening the box, part segmentation labels the entire lid of the box with the affordance 'openable'. However, to open the box correctly, humans leverage the top and bottom right corners of the box. Another problem with labeled affordances [5] is that the number of task affordances is limited by the number of affordance classes the annotator considers. Such labeling procedures thus limit the affordance data to the annotator trying to guess: "Which object parts exist, and what could they be used for?".

We argue that actionable affordances should not be labeled but should be *extracted* from observations of humans performing tasks, for example, from human videos. This has several advantages. Firstly, extracting affordances from human interaction with objects ensures that the "real" task-specific affordance regions are detected (Figure 1). Secondly, this makes affordance specification more natural since humans can often find it easier to *show* the object region to interact with rather than label and segment it correctly in an image. Moreover, the ability to extract the affordance region is very useful when learning from human demonstration or preferences i.e. when learning to interact with a novel object or adapt to end-users' needs. Another advantage is that, even with unsupervised videos of human interaction with objects, we only need to narrate or annotate what task is being done by the human in the video, which naturally gives us the affordance text label. The affordance data obtained from this procedure thus answers the questions: "What task is being performed with the object(s)?" and "Which precise object regions are being used?".

Egocentric videos of humans performing tasks are an attractive option for extracting affordance data [13, 14, 15, 16, 17] since they include object interactions which are close to and in the field of view of the camera. Recently, Goyal et al. [18] and Bahl et al. [19] have shown that videos from datasets such as EPIC kitchens [13] and Ego4D [15] can be used to succesfully segment regions of interest in objects using weak supervision from hand and object bounding-boxes. However, these works focus on segmenting task-agnostic 'hotspot' interaction regions of objects. In this work we focus on extracting precise actionable affordance regions while also using narrations of the human activity as affordance text labels. We propose a novel method to extract affordance masks using recent video-based hand mask segmentation and in-painting techniques. We use the EPIC Kitchens VISOR dataset [14] and extract 90k affordance masks from 38 videos from 25 different kitchens.

Moreover in this work, we consider the problem of *bimanual* affordance detection, since several household tasks require both hands to interact with one or two objects. Bimanual affordance prediction is especially more challenging since directly labeling such affordance masks for each hand is non-trivial. To the best of our knowledge, ours is the first method that extracts bimanual affordances from videos which we then use to train a model to predict task-specific affordance masks based on a text prompt. Moreover, our affordance extraction procedure also provides auxiliary information relevant to bimanual taxonomy [20], such as whether the task requires the use of two hands and if the actions by both hands should be symmetric or can they be asymmetric.

Our overall contributions are as follows:

- We propose a method to extract precise actionable affordance regions from human videos using inpainting and hand mask segmentation.
- We introduce a dataset called 2HANDS consisting of 90k images with extracted actionable affordance masks, narration-based labels and auxiliary bimanual taxonomy annotations.
- We qualitatively show favourable performance of our method against existing affordance prediction methods in predicting actionable affordance regions.
- To the best of our knowledge, we propose the first method to extract & predict bimanual affordance regions in RGB images.

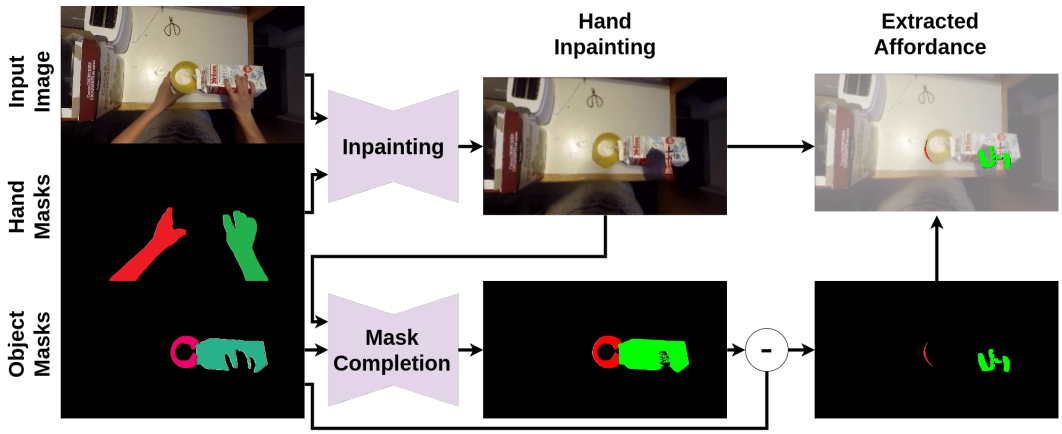

Figure 2: Affordance extraction pipeline. Given densely segmented sequences of the hands and objects (using XMem [21] with sparse segments), we use an inpainting model [22] to fill in the masked hand area. With the inpainted image and the object masks, we use [23] to "complete" the object masks where the hands have been masked out from. We can then extract the affordance for the given task as the difference between this completed mask and the original object mask.

2 Method

2.1 Affordance extraction

Our aim is to use videos of humans performing or demonstrating tasks to extract precise affordance masks, and use the task narration/annotation as the affordance text label. For the most part, this involves close inspection of the hands and object contact regions in the videos. Several recent methods [24, 25] have shown impressive performance in hand and object segmentation and reconstruction. However, the challenge in affordance region extraction lies in the fact that the hand typically occludes the object region that is interacting with. Bahl et al. [19] get around this problem by only considering videos where objects are initially un-occluded before the interaction, and only use the hand bounding box to denote the interaction region. However, not only is this a limiting assumption, but the bounding boxes can only be used to detect hotspots and do not provide *precise* object affordance regions. Thus, we propose a method that leverages recent advances in hand inpainting [22] and object mask completion [23] to extract precise affordance regions.

To demonstrate our method, we use egocentric videos from EPIC KITCHENS [13], which consists of $\sim$100 hours of human videos in kitchens. We use the VISOR [14] annotations of the dataset which contain sparse hand and object mask segmentations and a binary label denoting if the hand is in contact with the object or not. Note that we can also use other video datasets like Ego4D [15] or EgoExo4D [17] along with modern hand segmentation methods [24] to extract hand and object masks. Crucially, to get dense masks and obtain consistency over the video, we use a video object segmentation network [21] to propagate the hand and object masks. We then use hand inpainting [22] to inpaint the hand masks, followed by a mask completion stage [23] to complete the masks of the now un-occluded objects. The affordance region of the object that the hand was interacting with is now obtained by simply subtracting the hand-occluded object mask from the completed object mask. The full pipeline is shown in Figure 2. For bimanual affordances, it is also useful to classify the affordances into a bimanual taxonomy [20]. Thus we also track the frame-by-frame mean hand positions to annotate whether the task requires symmetric or asymmetric hand actions.

With the above procedure, we obtain a dataset of 90k images with extracted actionable affordance masks, narration-based labels, and auxiliary bimanual taxonomy annotations. We call this dataset 2HANDS, i.e., the 2-Handed Affordance + Narration DataSet.

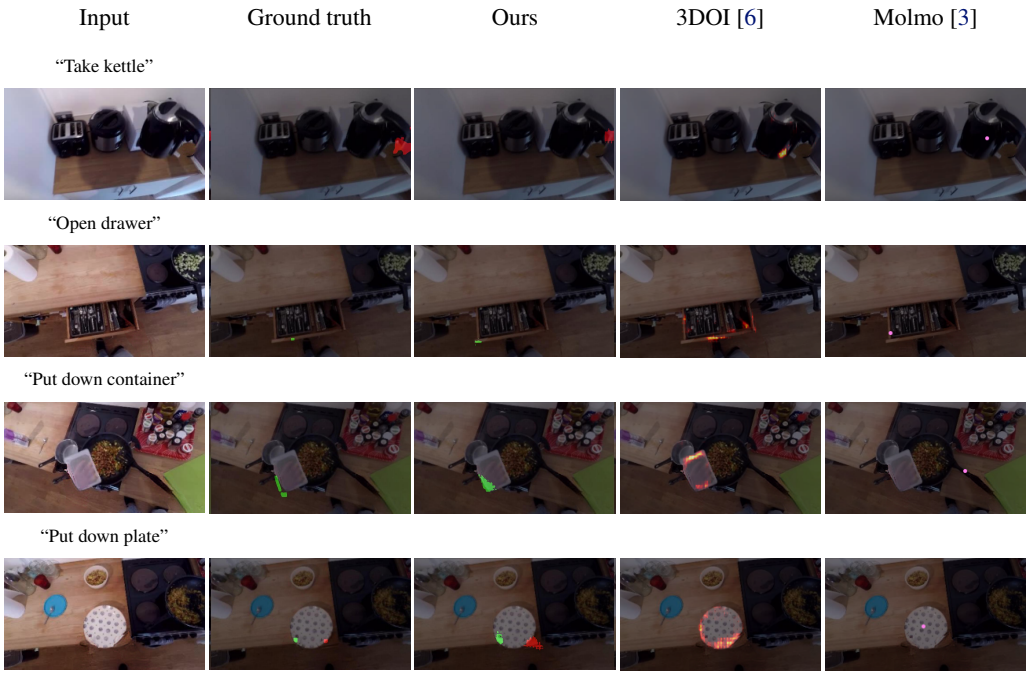

Figure 3: Qualitative results of our network 2HandedAff, compared against baselines. We see that the network can predict actionable regions and can often provide valid solutions even if they do not correspond exactly to the ground truth, for instance, in the "open drawer" example.

2.2 Affordance prediction

Our goal is to take an input image and an affordance text input, e.g. "pour tea from kettle", and segment the affordance region in the image. To enable the use of any natural language affordance text, we use a CLIP text encoder [1] to encode the text prompt. We encode the input image using a pre-trained segmentation backbone, SAM-Vit-B [26]. We concatenate the CLIP and SAM embeddings and pass them through two SAM-style mask decoders to predict the affordance masks for the left and the right hand. Since the affordances can be bimanual, eg. picking up a pot requires two hands or if two objects need to be handled simultaneously, we also classify the affordance masks into bimanual taxonomy classes, 'left-hand', 'right-hand', 'both hands - symmetric' and 'both hands - asymmetric' using a separate full-connected classifier decoder. We name our network '2HandedAff'.

The segmentation loss for the right and left mask prediction is a combination of the standard cross-entropy loss and the Dice loss [27]. The bimanual taxonomy loss is a standard cross-entropy loss. We can train a model to predict affordances based on the 2HANDS dataset. We split the data into 80k training, 10k validation, and 5k test images. We train for 100 epochs (32 hrs) on 8 A100 GPUs.

3 Preliminary experimental results

We show qualitative results of bimanual affordances predicted by our network on the test dataset of 2HANDS. Though no direct baseline exists to compare our method to since they do not consider bimanual affordances or actionable regions, we use two alternative types of methods as baselines: (i) 3DOI [6]: An affordance prediction baseline for two hands that uses labeled affordances on similar data. Since 3DOI needs query points on the image to work, we provide ground truth query points in the centers of the objects of interest; (ii) Molmo [3], a vision-language model that can point to different places in the image based on a free-form text prompt. Qualitatively, we find that our 2HandedAff network compares favorably to the baselines and is able to predict the actual affordance regions (Figure 3). Moreover, even when 2HandedAff deviates from the ground truth, it still segments reasonable solutions to perform the task.

4 Conclusion

In this work, we proposed 2HandedAfforder, a framework for extracting precise, meaningful and actionable bimanual affordances from human videos in unstructured environments. We have introduced a dataset of extracted actionable affordance regions from human videos using advances in video segmentation and inpainting. Our experiments demonstrate that 2HandedAfforder can predict meaningful task-oriented bimanual affordances compared to other works, thereby showcasing the effectiveness of our data extraction pipeline and proposed model. In the future, we plan to extend our method to other egocentric video datasets and to other tasks beyond the kitchen context. We also plan to use activity recognition methods to extract narrations for the videos automatically.

Acknowledgments

This research is funded by the European Union's Horizon program under grant agreement no. 101120823, project MANiBOT and the German Research Foundation (DFG) Emmy Noether Programme (CH 2676/1-1).

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, and et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[3] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, J. Lu, T. Anderson, E. Bransom, K. Ehsani, and H. N. et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models, 2024. URL https://arxiv.org/abs/2409.17146.

[4] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096, 2024.

[5] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435. IEEE, 2023.

[6] S. Qian and D. F. Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023.

[7] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024.

[8] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465, 2023.

[9] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[10] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[11] J. He, S. Yang, S. Yang, A. Kortylewski, X. Yuan, J.-N. Chen, S. Liu, C. Yang, Q. Yu, and A. Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[12] J. Lee, A. D. Tjahjadi, J. Kim, J. Yu, M. Park, J. Zhang, Y. Li, S. Kim, X. Liu, J. E. Froehlich, Y. Tian, and Y. Zhao. Cookar: Affordance augmentations in wearable ar to support kitchen tool interactions for people with low vision. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, 2024.

[13] D. Damen, H. Doughty, G. M. Farinella, , A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130: 33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

[14] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[15] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[16] L. Zhang, S. Zhou, S. Stent, and J. Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022.

[17] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[18] M. Goyal, S. Modi, R. Goyal, and S. 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.

[19] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[20] F. Krebs and T. Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters*, 7(4):11031–11038, 2022.

[21] H. K. Cheng and A. G. Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[22] M. Chang, A. Prakash, and S. Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] A. Sargsyan, S. Navasardyan, X. Xu, and H. Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7335–7345, 2023.

[24] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[25] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024.

[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[27] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.