FPHA-AFFORD: A DOMAIN-SPECIFIC BENCHMARK DATASET FOR OCCLUDED OBJECT AFFORDANCE ESTIMATION IN HUMAN-OBJECT-ROBOT INTERACTION

S.Muzamil Hussain.S[1] *Liu Liu*[2] *Wenqiang Xu*[1] *Cewu Lu*[1]

[1] Department of Computer Science, Shanghai Jiaotong University, P.R. China
[2] Institute of Intelligent Machines, University of Science and Technology of China, P.R. China

ABSTRACT

In human-object-robot interactions, the recent explosion of standard datasets has offered promising opportunities for deep learning techniques in understanding the functionalities of object parts. But most of existing datasets are only suitable for the applications where objects are non-occluded or isolated during interaction while occlusion is a common challenge in practical object affordance estimation task. In this paper, we attempt to address this issue by introducing a new benchmark dataset named FPHA-Afford that is built upon the popular dataset FPHA. In FPHA-Afford, we adopt egocentric-view to pre-process the videos from FPHA and select part of the frames that contain objects under the strong occlusion of hand. To transfer the domain of FPHA into object affordance estimation task, all of the frames are re-annotated with pixel-level affordance masks. In total, our FPHA-Afford collects 61 videos containing 4.3K frames with 6.55K annotated affordance masks belonging to 9 classes. Some of state-of-the-art semantic segmentation architectures are explored and evaluated over FPHA-Afford. We believe the scale, diversity and novelty of our FPHA-Afford could offer great opportunities to researchers in the computer vision community and beyond. Our dataset and experiment code will be made publicly available on https://github.com/Hussainflr/FPHA-Afford

***Index Terms*—** Human-Object-Robot Interaction, Object Affordance Estimation, Semantic Segmenation, FPHA-Afford

1. INTRODUCTION

Human-object-robot interaction is a classic research topic in the deep learning and robotics communities. Although tremendous researchers have shown a great success in object recognition task and developed applicable systems to precisely recognize or detect objects using computer vision techniques, only a few of them focus on finding a feasible way to predict how to interact with those object safely and appropriately. [1] In recent years, Gibson coined the term “Affordances”, which indicates “action possibilities” that could carry out with the objects (e.g a cup affords contain and grasp, a bottle affords wrap-grasp). Learning object affordances from images or videos is of great importance in robotics, along with accurate object recognition robots also require the ability of object affordance estimation and functionality understanding to interact with the recognized objects properly.

To learn object affordances in human-object-robot interactions, significant works have been done on learning visual affordance and functionalities understanding of objects and scenes[2]. In this case, these current methods focus on detecting and segmenting non-occluded and isolated object regions [3, 4, 5, 6], and the developed systems might work well in the applications without occlusion situations such as robotic pick and place. However, due to the difference between controlled and practical environments, these systems could achieve satisfied performance in predicting affordances for the occluded objects that might be more common in practical human-object-robot interaction systems. This could be attributed to the requirement of ideal data input. Most of existing methods are trained and validated on some specific datasets that collect images in a controlled environment with non-occluded annotated objects and action labels. Therefore, lack of appropriate datasets limits the advance of human-object-robot interaction applications.

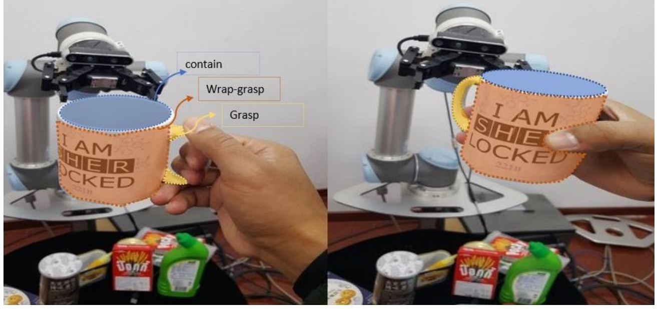
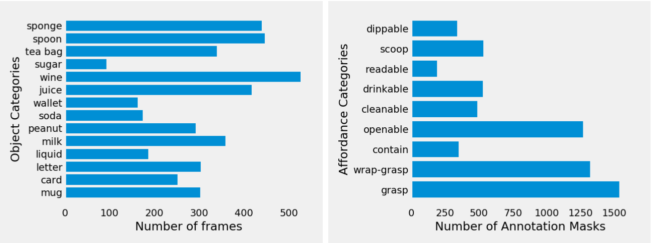


Fig. 1: FPHA-Afford is proposed to learn occluded object affordances in human-object-robot interaction.

In this paper, we attempt to address this issue by introducing a new domain-specific benchmark dataset that is built upon First-Person Hand Action (FPHA) so we named our dataset FPHA-Afford. To transfer the domain of original FPHA into our targeted occluded object affordance estimation task, we re-annotate the frames in FPHA with manual pixel-level annotations. Besides, following the strategy of [7], we pre-process all of the frames by adopting the egocentric view to make sure the video frames are considered as the view of agent camera. In total, our FPHA-Afford collects 61 videos containing 4.3K frames with 6.55K annotated affordance masks. In particular, in FPHA-Afford, 10 actions are being carried out with 14 different objects and 9 object affordances classes are manually annotated. Moreover, in our purposed FPHA-Afford dataset objects are under the strong occlusion of hand with different affordancess performed. We aim to find all object affordance regions in the frame where agent can take actions safely and appropriately. Several experiments are constructed over our FPHA-Afford using the state-of-the-art semantic segmentation techniques (e.g. FCN, delatedFCN [8, 9, 10]) to provide the baseline for occluded object

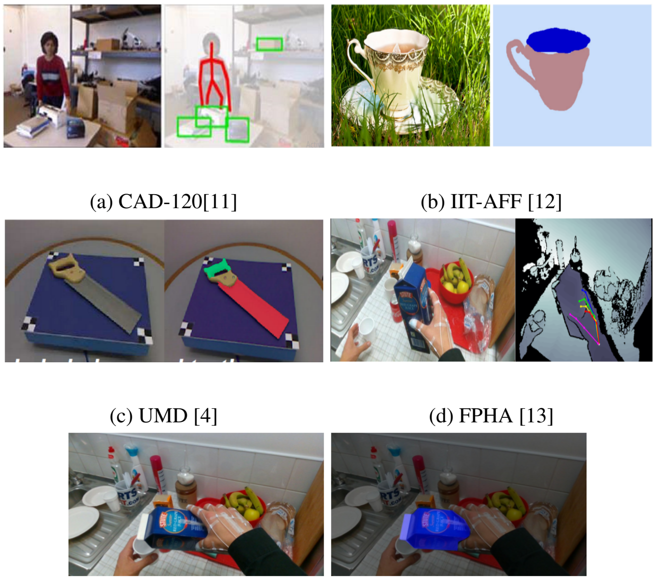
This work was supported in part by the National Natural Science Foundation of China under Grant 61772332 and in part by The National Key Technology R&D Program under Grant 2017YFA0700800 and 2017YFA0700801.



(a) Object frames distribution

(b) Affrodance masks distribution

Fig. 2: Statistics in our FPHA-Afford



(a) CAD-120[11]

(b) IIT-AFF [12]

(c) UMD [4]

(d) FPHA [13]

(e) FPHA-AFFORD **(Ours)**

Fig. 3: Visual comparison of different Datasets with FPHA-Afford

affordance estimation task.

In summary, our contributions in this paper are listed below:

- To the best of our knowledge, the first domain-specific dataset FPHA-Afford is published for object affordance estimation in human-object-robot interaction. Our FPHA-Afford contains 61 videos, 4.3K frames and 6.55K pixel-level affordance mask annotations. This benchmark will significantly promote the effectiveness and usefulness of applications of new human-object-robot interaction approaches.
- FPHA-Afford defines, analyzes and attempt to address a detailed domain-specific issue: occluded object affordance estimation, in which strong occlusion is a common problem in practical robot applications.
- We give comprehensive and depth performance evaluations of the state-of-the-art semantic segmentation techniques in FPHA-Afford. We believe that FPHA-Afford provides a feasible benchmark dataset and well facilitate further research on human-object-robot interaction field. Our dataset and code will be made publicly available.

Table 1: Comparison with some other popular datasets

	Pixel-level	Occlusion	Action
IIT-AFF [12]	✓		
UMD [4]	✓		
CAD-120 [11]		✓	✓
FPHA [13]		✓	✓
FPHA-Afford (Ours)	✓	✓	✓

2. RELATED WORK

Affordance learning and functionality understanding is of great importance in both computer vision and robotics fields. The concept 'affordance' is not limited to the objects but includes a variety of applications [affordancenet], from understanding human body parts [14] to environment affordances[15, 16]. In robotics, affordance is used to teach robots how to interact with real-world objects[4]. In a given environment, affordance could help anticipate possible actions since it represents all possible actions that can be carried out on or with objects. Besides, in the context of future action anticipation and prediction, affordance is also regarded as a specific signal to predict future human or robot actions [17, 11, 18]. In past, traditional image processing techniques are the major methods for affordance-based agent (e.g. a human or robot) activity recognition to recognize the interaction between agent and the surrounding environment [19, 20, 21]. For example, [7] proposed to learn social affordances to understand prohibited actions in the whole scene such as crossing road while the signal is red, happen in our daily life. Meanwhile, a similar method was also developed to detect object affordances from human-object interactions [22].

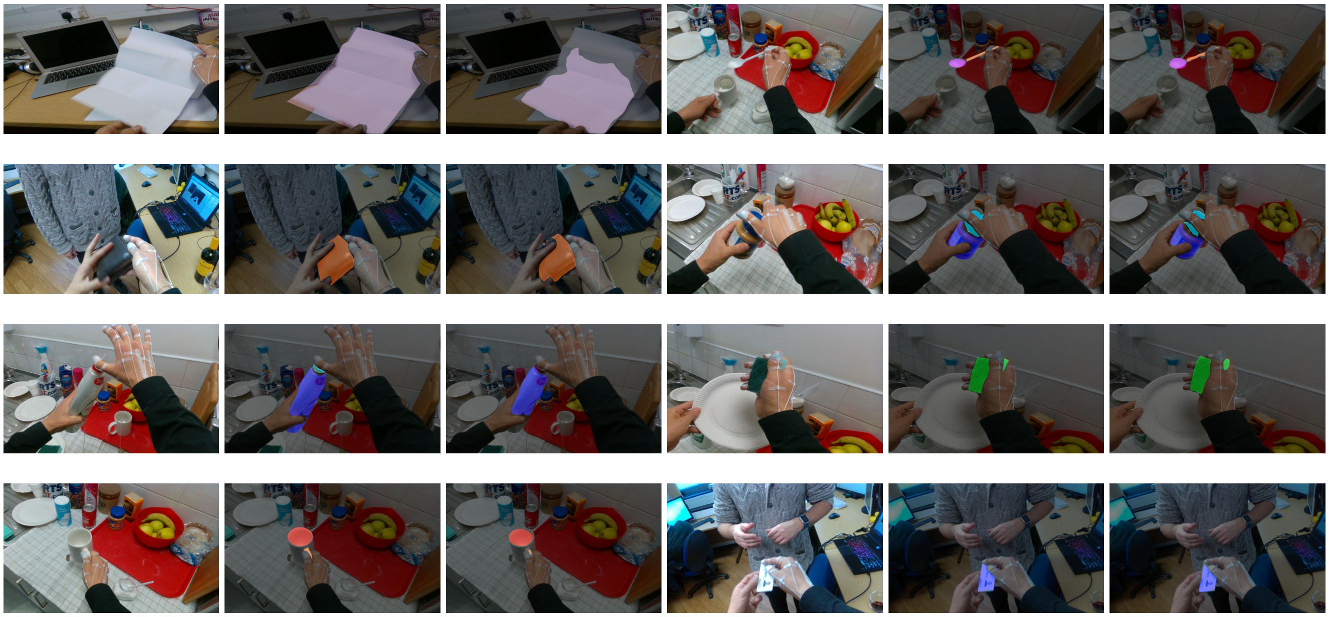
With the emergence of deep learning, a large number of researchers paid their attention on convolutional neural networks as main tool to design affordance detection frameworks. The work in [23] employed two deep neural networks to detect grasp affordances from RGB images. Apart from 3-channel images, RGB-D images were also considered as input for extracting deep features from neural networks to detect affordances [5]. In terms of network architecture, multi-scale convolutional neural network was designed by [15] localize environment affordances. In [24] a weakly supervised deep learning approach was adapted to segment object affordance in order to avoid highly expansive pixel ground-truth labels. Most recently, authors of [25] introduced to relationship-aware module along with Atrous Spatial Pyramid Pooling (ASPP) to improve affordance detection performance.

Despite that deep learning object affordances models have gained state-of-the-art improvement over hand-crafted geometric features[4], one limitation of these systems is that most of current datasets are difficult to be transferred into practical human-object-robot interaction applications since only non-occluded and isolated images considered, such as IIT-AFF[12] and UMD[4]. This results in a failure when objects are occluded or an action is being carried out. Keeping this gap in mind, we build a new domain-specific occluded object affordances dataset with pixel-level annotations specially designed for human-object-robot interaction scenarios.

3. FPHA-AFFORD DATASET

3.1. Data Collection

Instead of building a dataset from scratch that is laborious and time-consuming task, we decide to annotate FPHA [13] according to spe-



(a) Input (b) Ground-truth (c) Estimation (d) Input (e) Ground-truth (f) Estimation

🟪 Readable 🟧 Grasp 🟪 Scoop 🟦 Wrap-grasp 🟦 Openable 🟩 Cleanable 🟪 Drinkable 🟥 Contain

Fig. 4: Occluded Object Affordance Estimation Results on FPHA-Afford.

cific our requirements, which is publicly available RGB-D dataset for 3D hand-object interaction recognition that contains labels for 6D object pose, 3D hand pose, and action categories. Besides, FPHA is a large-scale dataset including 1,175 videos of 45 different activity categories performed by 6 subjects. In total, around 105K frames are annotated with accurate hand poses and action categories.

Motivated by FPHA that collects data in practical environment and the occlusion is considered in some of these images, we select part of videos from FPHA as the preliminary data in our dataset. After that, 61 videos are picked out that belong to 14 different objects and 10 actions. Then we manually filter a small part of frames in these videos to ensure that each frame (image) contains at least one occluded object. In this way, we finally obtain 4.3K frames to build our FPHA-Afford dataset.

3.2. Data Annotation

Due to our desire to achieve pixel-level object affordance estimation, we manually label every frame in our FPHA-Afford. With the help of the professional semantic segmentation annotation tool [26], we annotate all of the pixels per frame with 10 categories, in which 9 of them indicate object affordance classes such as grasp and the other one is negative label (background). To cover various affordances for the same object, we annotate every active object in a frame with a maximum of 3 affordance labels. For example, object cup might contain liquid or be grasped by hand simultaneously, so it could be labeled with 'contain' as well as 'grasp'. After annotation, in order to make sure that our FPHA-Afford is more suitable to be applied in robotics, we collect larger number of annotations of those affordance classes that occupy a small area in the whole frame and less quantity for those with large area. Finally, a total of 6.55K affordance masks are annotated in FPHA-Afford.

3.3. Pre-processing

In original FPHA [13] dataset, video frames have bit large field of view which is more prone to negative pixel leading to the unbalanced issue. In order to deal with unnecessary huge of amount of negative pixel and large field of view along with maintaining high frame resolution, we introduce dynamic cropping mechanism in FPHA-Afford that crops the frame into 1000×1000 by dilating global bounding box (contains all affordance segments).In the case that new dilated bounding box crossing frame boundary in any direction, these dilated bounding boxes are dynamically adjusted to opposite direction. In this way, we reduce some unnecessary negative pixel and maintain reasonable frame field of view.

3.4. Dataset Structure

For validating the performance of deep learning methods in FPHA-Afford, we split the whole dataset into training and testing subsets with nearly 1:1 ratio. In order to avoid overlap frames between these two subsets, the dataset is split at video level rather than frame level so we could make sure all the frames in training and testing set are from different videos. Specifically, there are 2.3K frames in training set and 2K for testing. The statistics of FPHA-Afford is shown in Fig 2. In particular, Fig 2(a) shows the number of frames per object category while Fig 2(b) shows that masks per affordance class. More details will be published with dataset. Comparing with other popular dataset such as CAD-120 [11], our FPHA-Afford is the first one that takes occlusion, pixel-level affordance and action labels into consideration at the same time. Fig 3 shows the visual comparison with different datasets and table 1 shows the superiority of proposed FPHA-Afford dataset over the existing datasets.

Table 2: Intersection over Union (IoU) Performance on FPHA-Afford test set

Method	Backbone	Affordance Categories									
		Dippable	Scoop	Readable	Drinkable	Cleanable	Openable	Contain	Wrap-Grasp	Grasp	mean
FCN [8]	VGG16	0.034	0.370	**0.355**	0.428	0.468	0.281	0.645	0.646	0.557	0.421
FCN [8]	ResNet50	**0.588**	0.729	0.317	0.815	0.793	0.539	0.816	0.821	0.691	0.679
DeepLab [9]	ResNet50	0.513	0.663	0.307	0.802	0.753	0.531	0.779	0.788	0.601	0.637
Encnet [27]	ResNet50	0.487	0.678	0.292	0.814	0.756	0.508	0.790	0.784	0.666	0.642
FastFCN [28]	ResNet50	0.563	**0.750**	0.339	**0.843**	**0.829**	**0.596**	**0.842**	**0.856**	**0.700**	**0.702**

4. EXPERIMENTS

4.1. Experimental settings

Following the standard design in computer vision, we consider one affordance for the region of pixels on the object that has the same functionality. In order to segment and classify different parts of object affordances under strong occlusion, we adopt semantic segmentation approaches to provide the baseline for FPHA-Afford. Therefore, four different state-of-the-art semantic segmentation methods are experimented to provide baseline including FCN[8], DeepLab[9], Encnet[27] and FastFCN[28]. Furthermore, in practical robotic systems, it is essential to consider computation complexity and memory footprint especially in object affordance estimation task. So, we choose Dilated ResNet50 as the feature extractor that could avoid memory overload and process images at a higher FPS compared with other heavy CNN backbones.

In our experiments, we use pixel-wise cross-entropy as loss function and follow the protocols presented in [28] during training. For fair comparison, we use the default training settings from [28]. Specifically, we set the initial learning rate to 0.001 and decrease it using ’poly’ strategy with power 0.9. Mini-batch Stochastic Gradient Descent (SGD) is employed as our optimization method with momentum 0.9 and weight decay 1e-4. Besides, We apply extensive data augmentation to avoid over-fitting, where the input frame is randomly scaled (0.5 to 2) and left-right flipped. Then these frames are cropped to 480×480 and grouped with batch size 16. The experiments are implemented using PyTorch and performed on 4 Titan-X Pascal GPUs (12GB memory per GPU). In terms of evaluation, Intersection over Union (IoU) is used as performance evaluation metrics following the standard of semantic segmentation field.

4.2. Result Discussion and Analysis

Table 2 the experimental results in testing set of FPHA-Afford under various methods. First of all, it is obvious that CNN backbone is one of the major component for the object affordance estimation where there appears a huge gap (near 1.5 times higher IoU) between FCN using VGG16 and ResNet50 as feature extractor in almost all the affordance classes. Comparing with other affordance categories, ’readable’ seems to be a difficult affordance to be segmented and recognized. This might be explained by the smaller training samples in FPHA-Afford and relatively larger mask would also result in lower IoU performance. Among these four state-of-the-art semantic segmentation architectures, FastFCN shows the best performance on our FPHA-Afford testing set with 0.731 mIoU obtained. This could be explained by the applied Joint Pyramid Upsampling (JPU) module that achieves effectively concatenating feature maps from various scales (Conv3-Conv5). Besides, in FastFCN, we also employ a multi-context module Pyramid Scene Parsing (PSP) to embed contextual features in difficult scenarios. So, even in some occluded objects or scenarios, the features of objects could also be maintained for precise affordance estimation.

4.3. Visualization

Finally, We demonstrate some object affordance estimation results in FPHA-Afford in Fig 4. Note that the results are estimated by FastFCN with ResNet50 backbone. As it could be seen, under strong occlusion problem in the objects, current methods could alleviate the effect and predict its corresponding affordance class. Furthermore, under the pixel-level affordance anntations, FPHA-Afford provides more detailed and finer object affordance information that much more significant for practical robot applications. This indicates that our FPHA-Afford could be applied to train a powerful model to be used in practical object affordance estimation systems. Therefore, Our proposed FPHA-Afford could well facilitate further research on human-object-robot interaction field.

4.4. Extensibility and Future Prospects

Even though FPHA-Afford provides defines and attempts to address a domain-specific issue occluded object affordance estimation by introducing a new benchmark and dataset, there exist several potential problems that point out future research directions. In order to make sure our FPHA-Afford could be applied human-object-robot interaction field, we are expected to capture more videos and images from practical environment as well as covering larger number of affordance classes. Moreover, we will also focus on developing a task-specific method accompanying with the dataset to further improve the affordance estimation performance.

5. CONCLUSION

In this paper, we highlight the problem of estimating object affordances in the context of human-object-robot interaction under strong occlusion. To address it, we collect a domain-specific benchmark dataset named FPH-Afford towards occluded object affordances estimation task. Our dataset is built upon FPHA and features pixel-level affordance annotation for every pixel of object of frames with 9 categories. We also implement and evaluate several state-of-the-art semantic segmentation approaches as baseline on FPHA-Afford. The experimental results demonstrate the usefulness and particularity of our FPHA-Afford dataset. We believe this work will help advance future research on human-object-robot interaction field.

6. REFERENCES

[1] James J Gibson, *The ecological approach to visual perception: classic edition*, Psychology Press, 2014.

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

[3] 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)*. IEEE, 2018, pp. 1–5.

[4] 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)*. IEEE, 2015, pp. 1374–1381.

[5] 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)*. IEEE, 2016, pp. 2765–2770.

[6] Krishneel Chaudhary, Kei Okada, Masayuki Inaba, and Xiangyu Chen, “Predicting part affordances of objects using two-stream fully convolutional network with multimodal inputs,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3096–3101.

[7] 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*, 2018, pp. 975–983.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] 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*, vol. 40, no. 4, pp. 834–848, 2017.

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[11] Hema S Koppula and Ashutosh Saxena, “Anticipating human activities using object affordances for reactive robotic response,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

[12] 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)*. IEEE, 2017, pp. 5908–5915.

[13] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim, “First-person hand action benchmark with rgb-d videos and 3d hand pose annotations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 409–419.

[14] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, “Refinenet: Multi-path refinement networks for high-resolution semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[15] Anirban Roy and Sinisa Todorovic, “A multi-scale cnn for affordance segmentation in rgb images,” in *European conference on computer vision*. Springer, 2016, pp. 186–201.

[16] Trung T Pham, Thanh-Toan Do, Niko Sünderhauf, and Ian Reid, “Scenecut: Joint geometric and object segmentation for indoor scenes,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.

[17] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena, “Learning human activities and object affordances from rgb-d videos,” *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[18] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena, “Structural-rnn: Deep learning on spatio-temporal graphs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5308–5317.

[19] Siyuan Qi, Siyuan Huang, Ping Wei, and Song-Chun Zhu, “Predicting human activities using stochastic grammar,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1164–1172.

[20] Jay Earley, “An efficient context-free parsing algorithm,” *Communications of the ACM*, vol. 13, no. 2, pp. 94–102, 1970.

[21] Tuan-Hung Vu, Catherine Olsson, Ivan Laptev, Aude Oliva, and Josef Sivic, “Predicting actions from static scenes,” in *European Conference on Computer Vision*. Springer, 2014, pp. 421–436.

[22] Hedvig Kjellström, Javier Romero, and Danica Kragić, “Visual object-action recognition: Inferring object affordances from human demonstration,” *Computer Vision and Image Understanding*, vol. 115, no. 1, pp. 81–90, 2011.

[23] Ian Lenz, Honglak Lee, and Ashutosh Saxena, “Deep learning for detecting robotic grasps,” *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[24] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall, “Weakly supervised affordance detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2795–2804.

[25] Xue Zhao, Yang Cao, and Yu Kang, “Object affordance detection with relationship-aware network,” *Neural Computing and Applications*, pp. 1–13.

[26] “http://www.robots.ox.ac.uk/ vgg/software/via/,” .

[27] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal, “Context encoding for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[28] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu, “Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation,” *arXiv preprint arXiv:1903.11816*, 2019.