One-Shot Affordance Detection

Hongchen Luo[1] , **Wei Zhai**[1,3*] , **Jing Zhang**[2†] , **Yang Cao**[1†] , **Dacheng Tao**[3]
[1] University of Science and Technology of China, China
[2] The University of Sydney, Australia
[3] JD Explore Academy, JD.com, China
{lhc12, wzhai056}@mail.ustc.edu.cn, jing.zhang1@sydney.edu.au, forrest@ustc.edu.cn, dacheng.tao@gmail.com

Abstract

Affordance detection refers to identifying the potential action possibilities of objects in an image, which is an important ability for robot perception and manipulation. To empower robots with this ability in unseen scenarios, we consider the challenging **one-shot affordance detection** problem in this paper, i.e., given a support image that depicts the action purpose, all objects in a scene with the common affordance should be detected. To this end, we devise a One-Shot Affordance Detection (OS-AD) network that firstly estimates the purpose and then transfers it to help detect the common affordance from all candidate images. Through collaboration learning, OS-AD can capture the common characteristics between objects having the same underlying affordance and learn a good adaptation capability for perceiving unseen affordances. Besides, we build a Purpose-driven Affordance Dataset (PAD) by collecting and labeling 4k images from 31 affordance and 72 object categories. Experimental results demonstrate the superiority of our model over previous representative ones in terms of both objective metrics and visual quality. The benchmark suite is at ProjectPage.

1 Introduction

The concept of affordance was proposed by the ecological psychologist Gibson [Gibson, 1977]. It describes how the inherent "value" and "meanings" of objects in an environment are directly perceived, and explains how this information can be linked to the action possibilities offered to an organism by the environment [Hassanin *et al.*, 2018]. After Gibson put forward the definition of affordance, a lot of cognitive psychologists made profound studies on the relationship between object affordance and perceiver intention in recent decades [Heft, 1989]. In particular, perceiving object affordance in unseen scenarios is a valuable capability and has a wide range of applications in scene understanding, action recognition, robot manipulation, and Human-Computer Interaction [Zhang and Tao, 2020].

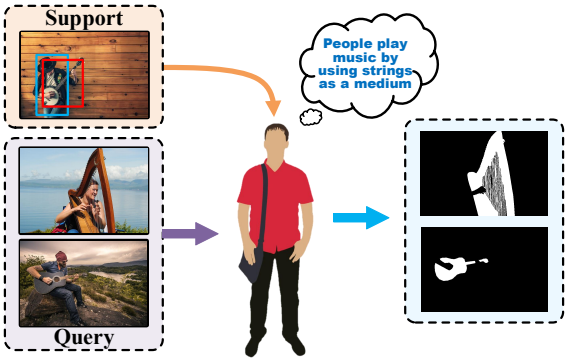

Figure 1: Illustration of perceiving affordance. Given a support image that depicts the action purpose, all objects in a scene with the common affordance could be detected.

To learn such capability of perceiving affordance, we consider the challenging **one-shot affordance detection** [1] problem in this paper, i.e., given a support image that depicts the action purpose, all objects in a scene with the common affordance should be detected (see Figure 1). Unlike the object detection/segmentation problem [Shaban *et al.*, 2017], affordance and semantic categories of objects are highly intercorrelated but do not imply each other. An object may have multiple affordances (see Figure 2), e.g., the sofa can be used to sit or lie down. Actually, the possible affordance depends on the person's purpose in real-world application scenarios. Directly learning the affordance from a single image without the guidance of purpose makes the model tend to focus on the statistically dominant affordances while ignoring other visual affordances that are possibly suitable for completing the task.

To address this problem: 1) We try to find clear hints about the action purpose (i.e., via the subject and object locations [Chen *et al.*, 2020]) from a single support image, which implicitly defines the object affordance and is a reasonable setting in real-world unseen scenarios. 2) We adopt collaboration learning to capture the inherent relationship between different objects to counteract the interference caused

*Wei Zhai is an intern at JD Explore Academy.
†Corresponding Author

[1] "Detection" follows the term pixel-wise detection task, which has also been used in the area of salient object detection.

Figure 2: The part (A) shows that objects usually have multiple affordance. The part (B) shows that the objects with different semantic categories may have the same affordance.

by visual appearance differences and to improve generalization. Specifically, we devise a novel **O**ne-**S**hot **A**ffordance **D**etection (**OS-AD**) network to solve the problem. Taking an image as support and a set of images (5 images in this paper) as a query, the network first captures the human-object interactions from the support image using a purpose learning module (PLM) to encode the action purpose. Then, a purpose transfer module (PTM) is devised to use the encoding of the action purpose to activate the features in query images that have the common affordance. Finally, a collaboration enhancement module (CEM) is devised to capture the intrinsic relationships between objects with the same affordance and suppress backgrounds that are irrelevant to the action purpose. In this way, OS-AD can learn a good adaptation capability for perceiving unseen affordances.

Moreover, the existing datasets still have gaps relative to real application scenarios, due to the limitation of its diversity. The affordance detection for scene understanding and general applications should be able to learn from the human-object interaction when the robot arrives at a new environment and retrieves the objects in the environment, rather than just finding objects with the same categories or similar appearance. To address the limitations of the datasets, we propose the **P**urpose-driven **A**ffordance **D**ataset (**PAD**), which contains 4,002 diverse images covering 31 affordance categories as well as 72 object categories from different scenes. Moreover, we trained several representative models to address the problem and compare them with our OS-AD model comprehensively in terms of both objective evaluation metrics and visual quality.

Contributions (1) We introduce a new one-shot affordance detection problem along with a benchmark to facilitate the research for empowering robots with the ability to perceive unseen affordances in real-world scenarios. (2) We propose a novel OS-AD network that can efficiently learn the action purpose and use it to detect the common affordance of all objects in a scene via collaboration learning, resulting in a good adaptation capability that can deal with unseen affordances. (3) Experiments on the proposed PAD benchmark demonstrate that OS-AD outperforms state-of-the-art models and can serve as a strong baseline for future research.

2 Related Work

2.1 Affordance Detection

The problem of visual affordance detection has been investigated for decades [Hassanin *et al.*, 2018], it is a widely used strategy in AI community to perceive action intentions and thereby to infer visual affordance from the image/video of human and objects. Early works mainly attempted to establish and learn an association between the apparent characteristics of objects and their affordance for perceiving affordance. [Myers *et al.*, 2015] proposed a framework for jointly locating and identifying the affordance of object parts and presented the RGB-D Part Affordance dataset, which is the first pixel-wise labeled affordance dataset. However, the affordances of objects do not simply correspond to representational characteristics, which are shifted in response to the state of interactions between objects and humans. Therefore, [Chuang *et al.*, 2018] considered the problem of affordance reasoning in the real world by taking into account both the physical world and the social norms imposed by the society, and constructed the ADE-Affordance dataset based on ADE20k [Zhou *et al.*, 2017]. Since the change of an object's affordance state is usually due to the interaction between the human and the object, further research has begun to consider human action as a cue for learning affordance. [Fang *et al.*, 2018] used human-object interactions in demonstration videos to predict the affordance regions of static objects via linking human actions to object affordance, and proposed the OPRA dataset for affordance reasoning. Different from the above existing works, our proposed method aims to explore the multiplicity of affordance by a collaborative learning strategy. To a certain extent, our work conforms to Gibson's definition of affordance that "it implies the complementarity of the animal and the environment". Since there exist multiple potential complementarities between animal and environment, it leads to multiple possibilities of particular affordance. To address this issue, we present a novel task of one-shot affordance detection, in which action intension is introduced through support image to alleviate the multiplicity of affordance.

2.2 One-Shot Learning

Recently, one-shot learning has received a lot of attention and substantial progress has been made based on metric learning using the siamese neural network [He *et al.*, 2021; Koch *et al.*, 2015; Snell *et al.*, 2017]. Besides, some works build upon meta-learning and generation models to achieve one-shot learning. Specifically, [Michaelis *et al.*, 2018] proposed the problem of one-shot segmentation in clutter, finding and segmenting a previously unseen object in a cluttered scene based on a single instruction example. [Zhu *et al.*, 2019] proposed a one-shot texture retrieval, given an example of a new reference texture, detecting and segmenting all the pixels of the same texture category within an arbitrary image.

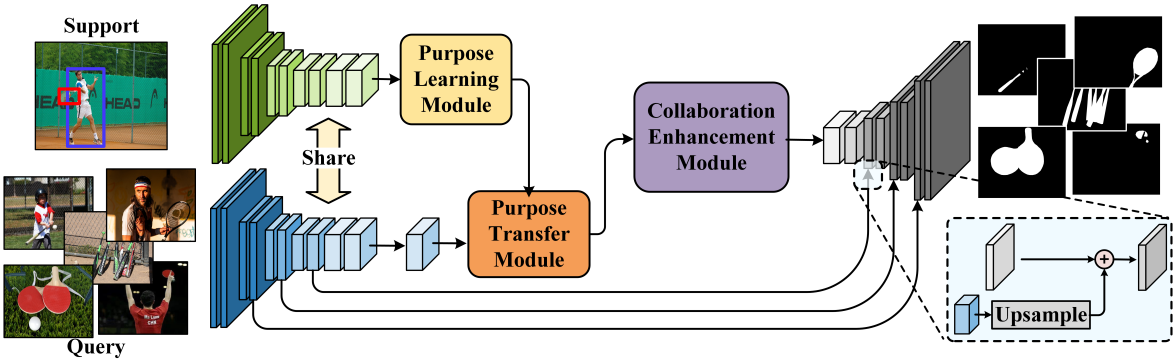

Figure 3: Our One-Shot Affordance Detection (OS-AD) network. OS-AD consists of three key modules: Purpose Learning Module (PLM), Purpose Transfer Module (PTM), and Collaboration Enhancement Module (CEM), which are detailed in Figure 4.

Different from these studies, we aim to identify the potential action possibilities (affordance) of objects in unseen scenarios, given a single support image that implicitly defines the action purpose without an explicit mask to denote the object affordance, which is very challenging but of practical value.

3 Method

Our One-Shot Affordance Detection (OS-AD) network is shown in Figure 3. During training, it receives a single support image that implicitly defines the affordance via human-object interactions, e.g., in the form of bounding boxes of the person and the object interacting with the person, which are weaker signals and easier to acquire than a pixel-wise affordance mask. Meanwhile, a set of query images (5 in this paper) containing objects with the same affordance. Our goal is to estimate the action purpose from the support image, transfer it to the query images, and learn to segment all the objects having the same affordance. To this end, three modules named PLM, PTM, and CEM are devised. During testing, OS-AD can recognize the common affordance from any number of query images given a single support image.

3.1 Framework

Given a set of query images $\mathcal{I} = \{I_1, ..., I_n\}$ and a support image I_{sup} containing human-object interactions, we first extract the features of $\mathcal{I}$ and I_{sup} using resnet50 [He *et al.*, 2016] to obtain their feature representations $\mathcal{X} = \{X_1, ..., X_n\}$ and X_{sup}, respectively. We then feed X_{sup} and the bounding boxes of the human and object into PLM to extract information about the human-object interaction and encode the action purpose. As shown in Figure 3, our network needs to estimate the purpose that the person wants to swing. Subsequently, we feed the feature representation of the action purpose and $\mathcal{X}$ into PTM, which transfers the action purpose to $\mathcal{X}$, enabling the network to learn to attend those objects with that affordance. Finally, the encoded features are fed into CEM, which captures the intrinsic connections between objects with the same affordance and suppresses irrelevant object regions, predicting the common affordance masks.

3.2 Purpose Learning Module

As shown in Figure 4 (a), we feed the X_{sup} and the bounding boxes of the person and object into the purpose learning module to estimate the action purpose of the person. We first use the bounding boxes to extract the features of the person and object from X_{sup} as X_{H} and X_{O}. Subsequently, inspired by human-object interaction and visual relationship detection [Zhan *et al.*, 2019; Zhan *et al.*, 2020], the features of the instance (person or object) can provide guidance information about where the network should focus, we use the features of the person and object to activate X_{sup} respectively to obtain M_{H} and M_{O}:

$$M_{\text{O}} = \text{Softmax}(f_{\text{O}} \otimes X_{\text{sup}}) \otimes X_{\text{sup}}, \quad (1)$$

$$M_{H} = \text{Softmax}(f_{\text{H}} \otimes X_{\text{sup}}) \otimes X_{\text{sup}}, \quad (2)$$

where f_{O} and f_{H} are the representations of X_{O} and X_{H} respectively after global maximum pooling (GMP). $\otimes$ denotes element-wise product. We then use f_{O} to guide the network to focus on the area of human-object interaction M_{HO}:

$$M_{\text{HO}} = \text{Conv}(f_{\text{O}} \otimes X_{\text{H}}). \quad (3)$$

Then, M_{HO} is multiplied with M_{H} and M_{O} respectively to activate the relevant features of human-object interaction on the global features, which are added together and go through a GMP layer to get the encoding of the action purpose F_{sup}:

$$F_{\text{sup}} = \text{MaxPooling}((M_{\text{HO}} \odot M_{\text{H}}) + (M_{\text{HO}} \odot M_{\text{O}})), \quad (4)$$

where $\odot$ denotes position-wise dot product.

3.3 Purpose Transfer Module

After obtaining F_{sup}, we transfer it to each query image to segment the objects that can fulfill that purpose, i.e., having the common affordance. As shown in Figure 4 (b), we perform correlation calculations between F_{sup} and each position of the query image. After normalization, we get the attention probability map that may contain objects with the common affordance. It is used to enhance the feature that may contain the affordance object and suppress the area that does not contain the affordance object. Finally, the attended feature is added back to the original feature to get $\mathcal{X}_{\text{T}}$:

$$X_{\text{T}_i} = X_i + \text{Softmax}(X_i \otimes F_{\text{sup}}) \otimes X_i, i \in [1, n]. \quad (5)$$

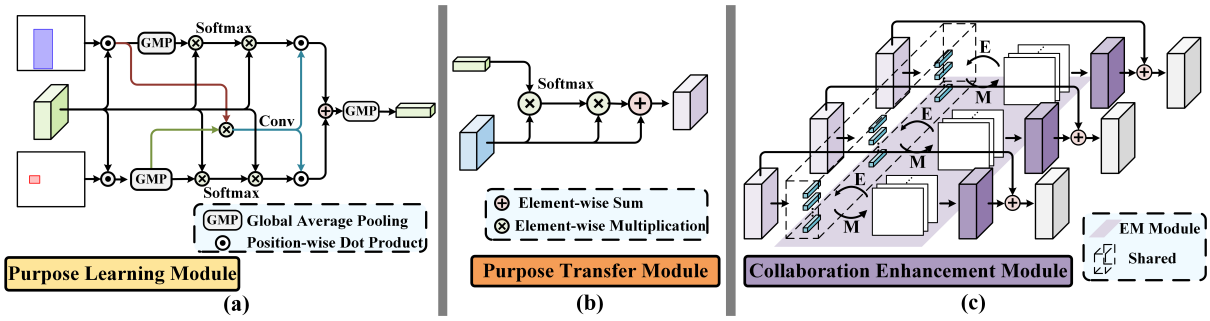

Figure 4: (a) PLM aims to estimate action purpose from the human-object interaction in the support image. (b) PTM transfers the action purpose to the query images via an attention mechanism to enhance the relevant features. (c) CEM captures the intrinsic characteristics between objects having the common affordance to learn a better affordance perceiving ability.

3.4 Collaboration Enhancement Module

There are some inherent characteristics between objects with the same affordance. For example, both cups and bowls have concave areas in the middle so they can be used to hold water. Thus, by discovering intrinsic features from the query image collection [Ma *et al.*, 2020], objects with the same affordance can be activated and unrelated regions can be suppressed, resulting in better segmentation results (see Figure 4 (c)).

Inspired by the Expectation-Maximization (E-M) [Dempster, 1977] algorithm and EMANet [Li *et al.*, 2019], we run the "E-step" and "M-step" alternately to obtain a compact set of bases, which are used to reconstruct the feature map of query images. During the iterative process, this set of bases can learn the common properties between the input set of features and thus can segment the common feature attributes. For a set of input feature maps $\mathcal{X}_{\mathrm{T}} = \{X_{\mathrm{T}_1}, ..., X_{\mathrm{T}_n}\}$ ($X_{T_i} \in R^{N \times C}$, $N = W \times H$), they firstly go through a convolution layer to get $\mathcal{F} = \{F_1, ..., F_n\}$. Then, a set of base $\mu \in R^{K \times C}$ is initialized. The E-step estimates the latent variable $\mathcal{Z} = \{Z_1, ..., Z_n\}$, $\mathcal{Z} \in R^{N \times K}$. The weight of the k-th basis for the j-th pixel of the i-th image is calculated as:

$$Z_{ijk} = \frac{\kappa(f_{ij}, \mu_k)}{\sum_{l=1}^{K} \kappa(f_{ij}, \mu_l)}, \tag{6}$$

where f_{ij} is the feature of the j-th position of the i-th image, κ denotes the exponential kernel function, i.e., $exp(,)$. Thus, we have $Z_i = Softmax(F_i \mu^T)$. The M-step updates the bases and calculates μ as the weighted average of $\mathcal{F}$:

$$\mu_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{L} z_{ijk} f_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{L} z_{ijk}}. \tag{7}$$

After convergence of the E-M iterations, we use μ and $\mathcal{Z}$ to reconstruct $\mathcal{X}$ to obtain $\tilde{\mathcal{F}}$: $\tilde{F}_i = Z_i \mu$. Finally, $\tilde{\mathcal{F}}$ is mapped to the residue space of $\mathcal{X}$ using convolution and added to $\mathcal{X}$ to obtain $\hat{\mathcal{X}}$, i.e., $\hat{X}_i = X_i + \mathrm{Conv}(\tilde{F}_i)$, $\hat{\mathcal{X}} = \{\hat{X}_1, ..., \hat{X}_n\}$.

3.5 Decoder

For the i-th image, the output of the $m - th$ layer is $P_i^m = \mathrm{Conv}(\mathrm{Unsample}(\mathrm{Conv}(X_i^m) + P_i^{m+1}), m \in [1, 4]$, where $P_i^5 = \mathrm{Conv}(\hat{X}_i^5)$. A convolutional prediction layer is used to get the final output, i.e., $D_i^m = \mathrm{Conv}(P_i^m), m \in [1, 5]$.

The cross-entropy loss is used as the training objective of our network. For the prediction D_i^m from the m-th layer of the i-th image, we calculate the loss $\mathcal{L}_i^m$ as:

$$\begin{aligned} \mathcal{L}_i^m = -\frac{1}{N}(& \sum_{j \in Y_+} logPr(y_{ij} = 1 | D_{ij}^m) \\ & + \sum_{j \in Y_-} logPr(y_{ij} = 0 | D_{ij}^m)), \end{aligned} \tag{8}$$

where $Pr(y_{ij} = 1 | D_{ij}^m)$ is the prediction map in which each pixel denotes the affordance confidence. Y_+ and Y_- denote the affordance region pixels set and non-affordance pixels set, $N = H \times W$. The final training objective is defined as:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{m=1}^{5} \mathcal{L}_i^m. \tag{9}$$

4 Experiments

4.1 Dataset

Data collection. We construct the Purpose-driven Affordance Dataset (PAD) with images mainly from ILSVRC [Russakovsky *et al.*, 2015], COCO [Lin *et al.*, 2014], etc. The affordance categories of the dataset is shown in Figure 5. Different images are collected according to the class keywords of the objects to make sure that the images in these datasets cover different scenes and appearance features of the objects, which constitute a challenging benchmark. Additionally, to make the dataset richer in categories, we obtain a part of the images from the Internet to expand the dataset. The description of each affordance and the object categories it contains are provided in the supplementary material. To benchmark different models comprehensively, we follow the k-fold evaluation protocol, where k is 3 in this paper. To this end, the dataset is divided into three parts with non-overlapped categories, where any two of them are used for training while the left part is used for testing. See the supplementary material for more details about the setting.

Data annotation. 1) Category labeling: We create a hierarchy for PAD dataset by selecting 72 common object categories (e.g. "cups", "bowls", "basketballs", etc.) and assigning affordance labels to each object category. An affor-

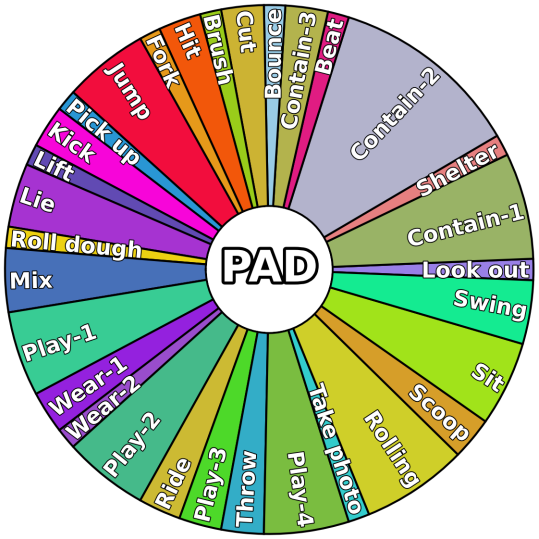

Figure 5: The classification system of the Purpose-driven Affordance Dataset (PAD), which contains 4,002 images covering 72 object classes and 31 affordance classes. See the supplementary material for more details of each affordance category.

dance category may contain multiple subcategories, e.g. objects with "Swing" affordance include "tennis rackets", "table tennis rackets", "golf clubs", "baseball bats", etc., and the appearance feature of these objects varies considerably. An object may contain more than one affordance category, e.g. "sofa" and "bench" have both "Sit" and "Lie" affordance labels. 2) Mask labeling: For the images from COCO etc., we select a portion of the mask labels from the original dataset. Since the above dataset may not annotate all the objects belonging to the same affordance, we select these images and annotate them manually. The images downloaded from the web are also manually annotated with the objects containing the affordance category. 3) Purpose image labeling: 656 purpose images are downloaded from the Internet which contains diverse human-object interactions. We annotate the bounding boxes of human and object accordingly.

4.2 Benchmark Setting

To provide a comprehensive evaluation, four widely used metrics are used to evaluate the performance of affordance segmentation. The **IoU** metric for segmentation task [Long *et al.*, 2015] is adopted in our task. **M**ean **A**bsolute **E**rror (**MAE**) [Perazzi *et al.*, 2012] is used to measure the absolute error between the prediction and ground truth (GT). **E-measure** (E_ϕ) [Fan *et al.*, 2018] is a metric that combines local pixels and image-level average values to jointly capture image-level statistics and local pixel matching information. Pearson **C**orrelation **C**oefficient (**CC**) [Le Meur *et al.*, 2007] is used to evaluate the correlation between the prediction and GT. We report the average metric score for all test images.

Our method is implemented in Pytorch and trained with the Adam optimizer [Kingma and Ba, 2014]. The backbone is resnet50 [He *et al.*, 2016]. The input is randomly clipped from 360×360 to 320×320 with random horizontal flipping.

	Metrices	i=1	i=2	i=3	Mean $\pm$ Std
UNet	IoU ↑	.186	.215	.226	$.209_{\pm.018}$
	E_ϕ ↑	.574	.558	.578	$.570_{\pm.009}$
	CC ↑	.338	.377	.344	$.353_{\pm.017}$
	MAE ↓	.162	.163	.169	$.165_{\pm.003}$
PSPNet	IoU ↑	.261	.244	.295	$.267_{\pm.021}$
	E_ϕ ↑	.640	.601	.636	$.626_{\pm.018}$
	CC ↑	.427	.409	.402	$.413_{\pm.011}$
	MAE ↓	.144	.142	.137	$.141_{\pm.003}$
CPD	IoU ↑	.258	.256	.317	$.277_{\pm.028}$
	E_ϕ ↑	.615	.601	.630	$.615_{\pm.012}$
	CC ↑	.413	.386	.433	$.411_{\pm.019}$
	MAE ↓	.123	.106	.132	$.120_{\pm.011}$
BASNet	IoU ↑	.239	.263	.281	$.261_{\pm.017}$
	E_ϕ ↑	.604	.598	.628	$.610_{\pm.013}$
	CC ↑	.310	.318	.339	$.322_{\pm.012}$
	MAE ↓	.130	.124	.146	$.133_{\pm.009}$
CSNet	IoU↑	.173	.210	.238	$.207_{\pm.027}$
	E_ϕ ↑	.557	.555	.557	$.556_{\pm.001}$
	CC ↑	.394	.392	.386	$.391_{\pm.003}$
	MAE ↓	.184	.162	.184	$.177_{\pm.010}$
CoEGNet	IoU ↑	.281	.262	.289	$.277_{\pm.017}$
	E_ϕ ↑	.674	.637	.645	$.652_{\pm.016}$
	CC ↑	.389	.350	.362	$.367_{\pm.016}$
	MAE ↓	.116	.110	.134	$.120_{\pm.010}$
Ours	IoU ↑	.401	.375	.407	$\mathbf{.394}_{\pm.011}$
	E_ϕ ↑	.732	.653	.687	$\mathbf{.691}_{\pm.032}$
	CC ↑	.540	.507	.501	$\mathbf{.519}_{\pm.017}$
	MAE ↓	.103	.116	.122	$\mathbf{.114}_{\pm.008}$

Table 1: Results of different methods on the PAD dataset under the 3-fold test setting. The best results are in **bold**.

We train the model for 40 epochs on a single NVIDIA 1080ti GPU with an initial learning rate $1e$-4. The number of bases in the collaboration enhancement module is set to K=256. The number of E-M iteration steps is 3. Besides, two segmentation models (**UNet** [Ronneberger *et al.*, 2015], **PSPNet** [Zhao *et al.*, 2017]), three saliency detection models (**CPD** [Wu *et al.*, 2019], **BASNet** [Qin *et al.*, 2019], **CSNet** [Gao *et al.*, 2020]) and one co-saliency detection models (**CoEGNet** [Fan *et al.*, 2021]) are chosen for comparison.

4.3 Quantitative and Qualitative Comparisons

The results are shown in Table 1, our results outperform all methods on all metrics in the three one-shot learning settings. Measured by the mean IoU value of the three-fold test, our model improves by 42.2% compared to the co-saliency detection approach. Compared to the best saliency detection model and the best segmentation model, our method improves by 42.2% and 47.6% respectively. This indicates that our method can effectively use the action purpose extracted from the support image to guide the segmentation of query images, rather than simply constructing a link between apparent features and affordance. The prediction results generated by each model are shown in Figure 6. It can be seen that our method can detect all the objects with common affordance according to the human purpose. In the 1st and 3rd rows of "Kick", the soccer ball and the punching bag are completely different in terms of apparent features, but both are kickable objects. This demon-

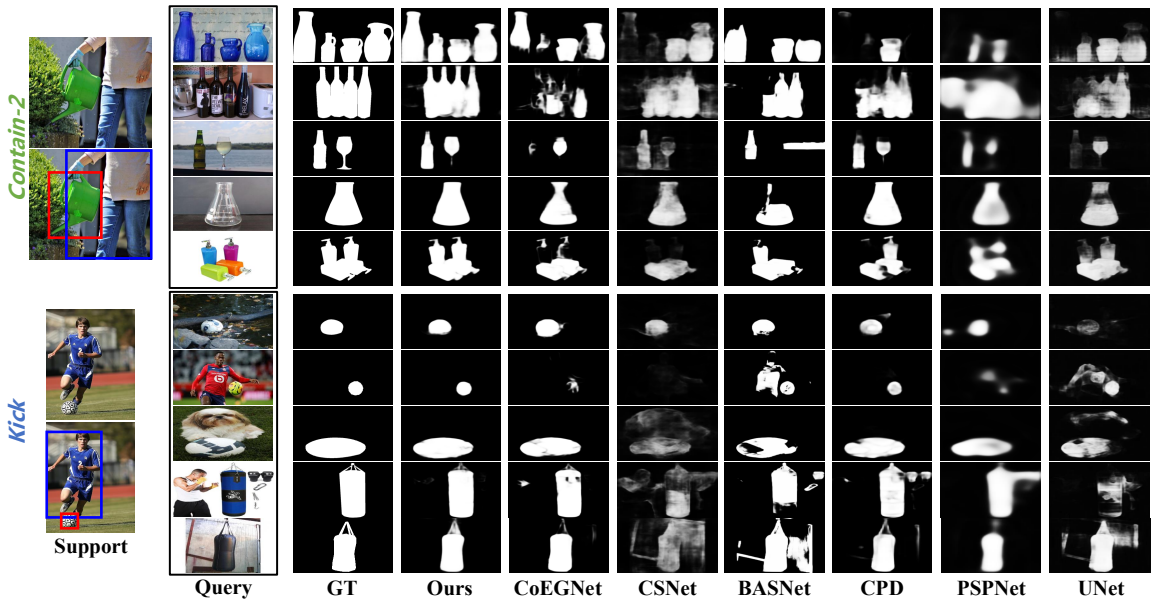

Figure 6: Visual results of different segmentation, saliency detection, co-saliency detection models and our OS-AD on the PAD dataset. OS-AD can learn a better capability to perceive the affordance of objects, i.e., segmenting all objects that complete this purpose and suppressing object regions that are not related to affordance. "Container-2" refers to the affordance category that objects can fill liquid.

	Kick	Play-4	Contain-2	Brush	Jump
UNet	.256	.332	.253	.082	.065
PSPNet	.450	.452	.283	.108	.122
CPD	.500	.505	.347	.140	.162
BASNet	.358	.408	.328	.114	.109
CSNet	.260	.393	.243	.084	.043
CoEGNet	.458	.453	.321	.107	.132
Ours	**.615**	**.593**	**.414**	**.246**	**.206**

Table 2: Comparison of part results on different affordance categories, using IoU as a metric. Where "Container-2" refers to the object with the availability of liquid and "Play-4" refers to the objects that a person blows through his mouth to make a sound. See the supplementary material for more details.

strates that our method can detect object affordance in the unseen scenarios by extracting the action purpose and collaborative learning strategy. Meanwhile, we calculate the IoU for each affordance class, and Table 2 shows the results for several of them. The results from the first three columns show that our method can detect object affordance with different apparent features but belonging to the same affordance, indicating that our model can capture the common relation of the objects in the unseen scenarios well and detect them more completely. The two categories with the worst results out of all the results are "Jump" and "Brush", respectively. The possible reasons for the poor performance are as follows: while in the "Brush" category, its accompanying actions are not obvious and the size of the objects are relatively small, making it more difficult for the network to extract the features of the objects. In summary, our approach can learn object affordances well by transferring human purposes to new objects and capturing common features between objects using collaborative learning strategies, with good generalization ability on the task of detecting object affordance in unseen scenarios.

5 Conclusion

In this paper, we make the first attempt to deal with a challenging task named one-shot affordance detection, which has practical meaning for real-world applications, such as empowering robots with the ability to perceive unseen affordance. Specifically, we devise a novel one-shot affordance detection (OS-AD) network and construct a benchmark named purpose-driven affordance dataset (PAD). OS-AD exceeds representative models adapted from related areas such as segmentation and saliency detection, which can serve as a strong baseline for this task. PAD contains 4k images with diverse affordance and object categories from different scenes, which can serve as a pioneer test bed for this task. In the future, we plan to deploy the model in a real-world robot to complete some well-defined tasks and evaluate its performance. Going a step further, we attempt to make such a test suite available to the community via a web-based interface.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105701, the National Natural Science Foundation of China (NSFC) under Grants 61872327, the Fundamental Research Funds for the Central Universities under Grant WK2380000001.

References

[Chen *et al.*, 2020] Zhe Chen, Jing Zhang, and Dacheng Tao. Recursive context routing for object detection. *IJCV*, 2020.

[Chuang *et al.*, 2018] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018.

[Dempster, 1977] A. P. Dempster. Maximum likelihood from incomplete data via the em algorithm. *JRSS*, 1977.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018.

[Fan *et al.*, 2021] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *TPAMI*, 2021.

[Fang *et al.*, 2018] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018.

[Gao *et al.*, 2020] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In *ECCV*, 2020.

[Gibson, 1977] James J Gibson. The theory of affordances. *Hilldale*, 1977.

[Hassanin *et al.*, 2018] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *arXiv*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[He *et al.*, 2021] Haoyu He, Jing Zhang, Bhavani Thuraisingham, and Dacheng Tao. Progressive one-shot human parsing. *AAAI*, 2021.

[Heft, 1989] Harry Heft. Affordances and the body: An intentional analysis of gibson's ecological approach to visual perception. *J Theory Soc Behav*, 1989.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[Koch *et al.*, 2015] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML workshop*, 2015.

[Le Meur *et al.*, 2007] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision research*, 2007.

[Li *et al.*, 2019] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[Ma *et al.*, 2020] Benteng Ma, Jing Zhang, Yong Xia, and Dacheng Tao. Auto learning attention. *NeurIPS*, 2020.

[Michaelis *et al.*, 2018] Claudio Michaelis, Matthias Bethge, and Alexander S Ecker. One-shot segmentation in clutter. *arXiv*, 2018.

[Myers *et al.*, 2015] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015.

[Perazzi *et al.*, 2012] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[Shaban *et al.*, 2017] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv*, 2017.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[Wu *et al.*, 2019] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[Zhan *et al.*, 2019] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *CVPR*, 2019.

[Zhan *et al.*, 2020] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. Multi-task compositional network for visual relationship detection. *IJCV*, 2020.

[Zhang and Tao, 2020] Jing Zhang and Dacheng Tao. Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 2020.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[Zhu *et al.*, 2019] Kai Zhu, Wei Zhai, Zheng-Jun Zha, and Yang Cao. One-shot texture retrieval with global context metric. *IJCAI*, 2019.