Perception and Segmentation of Human Hand Affordance on Everyday Objects

Sharon Chen Yongkyun Lee Sanghyun Yi

California Insitute of Technology
{schen5, ylee2, syi}@caltech.edu

Abstract

Gibson's affordance theory provided an ecologically appealing framework for developing robots. However, previous approaches that overly restrict the affordance or object domain, as well as use high-cost pixel-by-pixel annotation of datasets have limitations in implementing the many-to-many mapping between object and action/effect, i.e. the affordance. In this study, by focusing on hand actions, we propose a convolutional neural network (CNN)-based model that can both predict the hand action affordance of objects as a many-to-many mapping and segment objects by their hand actions through a weakly-supervised method. This model for predicting and segmenting human hand action affordances will provide a cost-effective method for detecting hand action affordances, and propel the development of robotic manipulation technologies, and ergonomic design of products.

1. Introduction

Gibson's affordance theory suggests that one's environment is perceived not simply in terms of its identity and spatial configuration but also more actively in terms of the possible actions that it suggests [5, 9]. The concept of affordance has been adopted for developing reactive and behavior-based robots that assume that robot actions are closely linked to perception; a robot's actions are executed based on a direct mapping of sensory input to the affordances for specific motor actions [14].

However, studies on affordance in robotics have two pitfalls. First, because the concept of affordance is a many-to-many mapping between objects and actions, teaching robots to learn the affordances of different actions on objects has been a challenging problem that has been studied on restricted domains or datasets. For example, studies have focused only on the learning of the affordances of a small number of action types (e.g., one of lift-ability, prehinsility, push-ability, etc.) or on the the affordances on a limited set of objects [2, 7, 8, 11, 18, 23, 25, 26].

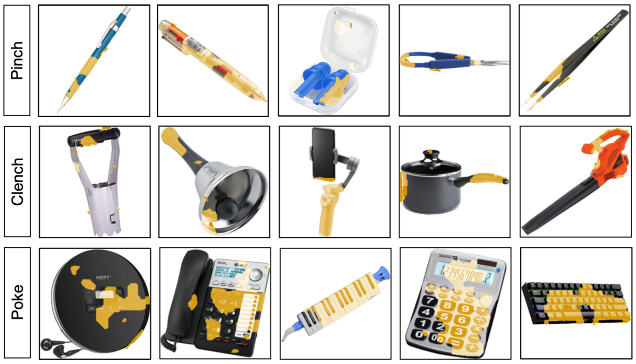


Figure 1. Examples of the proposed model's prediction and segmentation results. For instance, the first row images represent pinch-affording objects predicted by the model, while the yellow segments represent the corresponding pinch-conferring segments. We observed that our method of weakly-supervised segmentation produces reasonable results.

Second, even studies that involve detecting various affordance classes assume that an object, or each pixel of an image, corresponds to a single type of affordance. Not only is this inconsistent with the many-to-many nature of affordance, but it also could constrain the planning of complex behaviors. For example, while the blade of a kitchen knife has the *cut* affordance, the side of the blade can be used to *cruch* ingredients, by *pushing* it against them. However, simply labeling the ground truth affordance of the blade as *cut* eliminates the possibility of the latter usage. Therefore, it is critical to preserve the many-to-many characteristics when modeling affordance for robotic manipulation.

In this study, by focusing only on hand actions, we propose a CNN-based model that can both predict hand action affordance of objects as a many-to-many mapping and also segment objects based on hand actions in a weakly-supervised manner (Fig. 1). This prediction and segmentation model will provide a cost-effective method for detecting hand action affordances, which will propel the development of robotic manipulation technologies and ergonomic product design.

Novelty and Contributions Based on the findings from psychology/neuroscience that the majority of human hand actions during initial contacts with objects can be categorized into 4 actions (pinch, clench, poke and palm), we focused on how plausible it was to perform each of these actions on various objects [4, 18, 24, 27]. In this project we did not model the possible *effects* of objects, i.e. what the objects can do (e.g., cut, contain, display, etc.); we focused only on what can be done on the object, which may be a limitation of this study. However, given the vastness of possible effects categories, confining the domain to the perception of possible action on objects is a reasonable first step.

Second, considering the limitations of the datasets from previous literature on affordance, we create a new dataset of tool images from an online shopping mall. Unlike previous datasets for affordance detection, which include depth information and pixel-by-pixel annotations of affordances, our dataset is RGB-only and labeled with plausibility scores of all 4 actions for each tool globally, significantly reducing the cost of data collection.

Lastly, we developed a weakly-supervised method for performing affordance segmentation on objects by using our trained affordance score prediction model. We formulated the affordance score prediction problem as a multivariate regression that predicts the plausibility scores of each hand action based on an input tool image, allowing the many-to-many mapping of actions and objects. This turns out to be useful for pixel-by-pixel affordance segmentation by using gradient-based saliency maps. Moreover, this model allows the assignment of multiple affordances to a single pixel. We were able to generate reasonable segmentations for each hand action which were verified qualitatively and quantitatively.

2. Related Works

Human Object Affordance and Grasp Detection Human object affordance and grasp detection are the foundation for developing robotic manipulation. Early work in psychology and neuroscience demonstrated that human hand configurations on objects could be predicted using information about the object's size and convexity [3, 4, 18, 24]. Furthermore, it has been shown that hand postures associated with everyday tools are supported by structural and functional information about the tools [3, 16].

Researchers in robotics and computer vision have developed algorithms that mimic human-object interaction through data-driven approaches [2, 14]. For example, Saxena et al.'s work on robotic grasping showed that a feature-based statistical model trained on stereo of synthetic 2D images could generalize its learned information to grasp a variety of everyday objects [29]. Large image datasets of RGB-D images of tools have also been introduced in which each pixel is labeled with a specific affordance (cut-ability, grasp-ability) [15, 23, 26]. These large image corpora have enabled a deep learning-based approach to affordance detection that has achieved high levels of performance [7, 20, 26].

However, the fact that these past studies typically only use a limited class of action affordances limits the application of the algorithm only to tools that confer those few actions. Also, while the data are large enough to train deep neural networks, there are typically < 20 object categories in the datasets, which makes them quite limited [15, 23, 26].

Weakly-supervised segmentation Due to the inefficiency of annotating pixel-level ground truth labels for segmentation tasks, various approaches to developing weakly-supervised methods for segmentation have been investigated. Early work in object classification using CNNs have demonstrated that a neural network trained on object recognition could be used for foreground segmentation [30]. Various types of weak labels have been studied in the context of weakly-supervised semantic segmentation, and previous work utilizing image-level classification labels are closely-related to our work [17, 21, 32]. Past work have relied on CNN visualization techniques using gradient-based saliency maps such as the Class Activation Map (CAM) and explored the method of network training to refine CAM for improved segmentation performance [1, 32, 33].

3. Overall Framework

In this section, we describe our approach for weakly-supervised hand affordance segmentation using our hand affordance prediction network (Fig. 2).

We trained our affordance prediction network using global affordance labels for each object (Section 4). Because an object's hand action affordance can vary across its surface (e.g., the periphery of a smartphone can afford clenching while a button on it affords poking), we assumed that global-level labels for a specific hand affordance could be modeled as a continuous probability distribution, rather than a binary label indicating the existence of the affordance. We consider pinch, clench, poke, and palm as the possible affordance classes for each object, so we model each object as having 4 global probability distributions for hand action scores. As a result, our model predicts summary statistics about the distributions, or means, of each hand action distribution.

We used a CNN-based regression model to predict the mean scores of hand action affordance (Section 5.1). The model takes an RGB image of an object and predicts 4 scores: the mean scores of pinch, clench, poke, and palm. In addition, the affordance score prediction model predicts the size information that we labeled for each image (Section 4). This is because the input images do not include information about the size of the objects, although information about object size is essential for making accurate predictions of

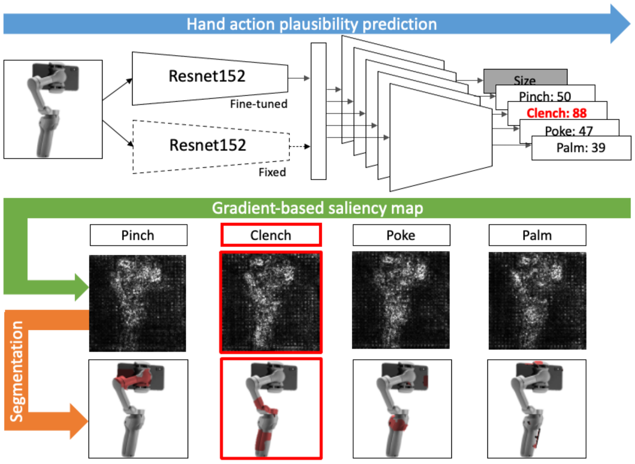


Figure 2. A schematic overview of the proposed method. First, we developed a model that predicted plausibilities of hand actions to a given object. Then, for each action, we determined the contribution of each pixel's unit change to the change in the plausibility predictions to create saliency maps. Finally, using those saliency maps, segmentations of the object conferring each hand action were obtained. In the example shown, the model predicts that the smartphone gimbal most highly affords *clench* and the clench segment effectively localizes the grip.

hand action affordance.

After training the model, we generated four saliency maps, each of which corresponds to an action type, by calculating gradient of scores for each pixel using guided backpropagation [31]. Each pixel value of the saliency maps indicates the network output's sensitivity to a unit change of the pixel in the original image. We used these saliency maps as inputs to a post-processing algorithm that outputs the segmentation of hand affordances on objects (Section 6). Note that there is no pixel-by-pixel supervised learning in the post-processing algorithm.

4. Dataset

A new image dataset of everyday objects was built for this project (Fig. 3). We scraped images of approximately 900 randomly-selected unique everyday objects from an online shopping mall, Amazon, for a total of 1000 images. These images all had white backgrounds, and there were multiple images for some objects viewed from different perspectives to allow for modeling of the affordance differences between viewpoints on objects (Fig. 3b).

We crowdsourced the labels for each image by asking Amazon Mechanical Turk (Mturk) users simple questions such as "How suitable is it to pinch to the shown object?" and displaying a slider for them to respond. The slider ranged from "Very unsuitable" to "Very suitable", which was later mapped to 0 to 100 scale with 100 being the most suitable. Each Mturker was asked to provide a response for each of 4 actions for each image. Prior to annotating images, participants were required to watch demonstration videos of each hand action and complete related quizzes so that the annotators understood the meaning of each hand action. Additionally, we collected familiarity scores for the displayed object 4 times for each image. We also included 4 catch questions in the annotation job for quality control. Data from annotators who failed to pass more than 2 catch questions were removed. As a result, we used data from 160 out of the total 227 Mturkers. There were 7.2 annotators per image on average. In other words, for each image, we have 7.2 annotations for each hand action type, and 4×7.2 familiarity annotations. Then, to get the ground truth for each action's affordance scores and familiarity for each image, we used the averages of each action's annotated scores and familiarity scores (Fig. 3a).

When we presented object images to annotators, we also displayed each image with a human silhouette to help the annotator understand the object's size relative to the human body (Fig. 3c). We manually determined the size of objects based on our knowledge of each object. Then, we used the number of pixels that the object was composed of, as well as the width and height of the object on the human silhouette image in terms of pixels, as the size information of each object.

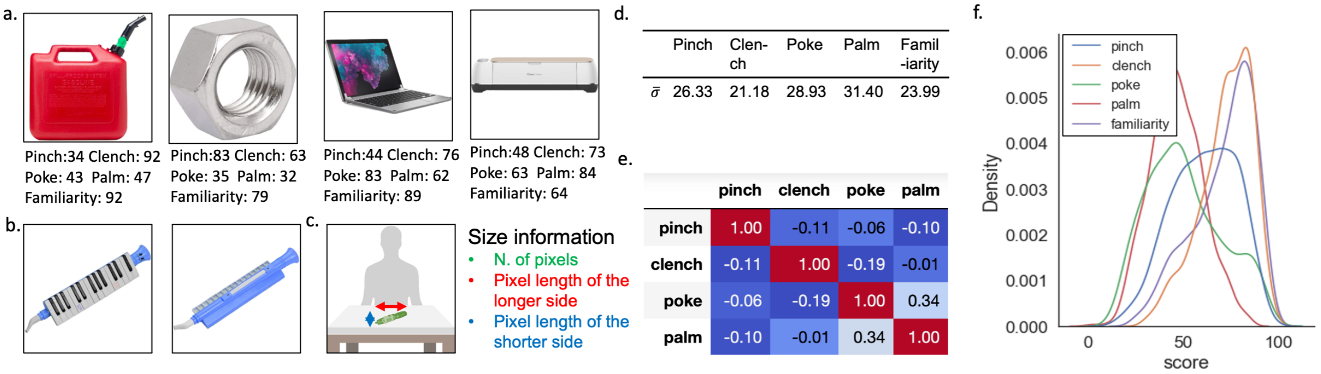
We would like to note that the average standard deviation of palm annotations (31.40) was larger than other action classes. Also, each image's average palm scores and poke scores are positively correlated (0.34). The average scores for each hand action and familiarity varied (For details, see Fig. 3d, e and f).

5. Affordance Score Prediction

5.1. Network Architecture

We adopted the output design choice from an aesthetic prediction study that aimed at predicting the distributions of liking scores on images [22, 13]. In this architecture, the network makes a 10-dimensional softmax output, from which the weighted sum produces a scalar value prediction. Therefore, we can interpret each element of the softmax output as a single bin of the binned distribution of the dependent variable. Adopting this architecture, our model calculates the weighted sum of a 10-dimensional softmax output for each hand action score, which each ranges from 0 to 100.

Different versions of our model have either one or two CNN feature extractors which were trained on Imagenet classification task [6]. In either version, only one CNN pathway was fine-tuned during training. In the two-pathway version, outputs of each CNN feature extractor were concatenated into a single vector. We made this de-



	Pinch	Clen-ch	Poke	Palm	Famil-iarity
$\bar{\sigma}$	26.33	21.18	28.93	31.40	23.99

	pinch	clench	poke	palm
pinch	1.00	-0.11	-0.06	-0.10
clench	-0.11	1.00	-0.19	-0.01
poke	-0.06	-0.19	1.00	0.34
palm	-0.10	-0.01	0.34	1.00

Figure 3. Example images and statistics of our newly collected dataset. **(a)** Example images and their annotations. **(b)** Example images taken from an item from different perspectives. **(c)** An overview of the definition of size information. When we presented objects to annotators, we also showed images with human silhouette to illustrate the object's size. **(d)** Each action's annotation variances averaged across images. **(e)** Correlation between average affordance scores for each pair of actions. **(f)** Average affordance score distributions for each action. For more information, see Section 4.

sign choice to preserve the object identity information from the Imagenet-pretrained network because knowledge about an object is an important factor in object manipulation [16].

The vector representation of an input image is then processed by 5 read-out fully-connected networks (FCNN), which make predictions for pinch, clench, poke, and palm scores and size information (number of pixels, pixel lengths of the longer and shorter side of the object on the human silhouette image. See Fig. 3c and Section 4). FCNNs for hand action scores have one 4096-dimensional hidden layer and make 10-dimensional softmax outputs with their weighted sums computed as predictions of action scores from 0 to 100. The layers of the networks were interleaved with ReLU non-linearities and dropout layers with dropout probability 0.5. The FCNN for size information has no hidden layer but makes a 3-dimensional output where the dimensions correspond to the three types of size information.

5.2. Experiments

We trained the network architecture described in Section 5.1 using Huber loss, which is robust against outliers [12], with a scaling factor of 1. The total loss function was a summation of the Huber loss for each action and size information prediction.

We used 5-fold cross-validation to evaluate our models. In each fold, 200 images were held out for evaluation, 600 images were used for training and remaining 200 images were used for early stopping to avoid overfitting. Note that objects from same categories can be both in the training and evaluation set, e.g. scissors can be found in both the training and the test set. We used random horizontal and vertical flips, color jittering, and random gray scaling for data augmentation.

Stochastic gradient descent was used with a momentum factor of 0.9. The learning rate decayed every 30 epochs from 0.0001 by the factor of 10. L2 regularization was used with a weight of 0.0005. When the best validation loss had not been updated for 35 epochs, the training was stopped and the weight checkpoint that updated the validation loss at the end was used as our final model for evaluation. Python 3.7, PyTorch 1.8.1, CUDA 10.2 and a single Nvidia 1080ti GPU were used throughout the experiments.

Model	RMSE	Correlation	Accuracy(%)
Res152+2path+size (R2s)	**14.80**	**0.448**	**64.35**
R2s, no repeated objects	14.90	0.415	62.57
Res152+2path	14.80	0.420	63.72
Res152+size	15.04	0.429	64.22
Res152	15.22	0.420	63.67
Res50	15.11	0.418	63.35
Res18	15.41	0.372	62.03
VGG19	14.99	0.415	62.65
VGG11	15.21	0.390	61.65
AlexNet	15.43	0.343	60.65

Table 1. 5-fold cross-validation evaluations of different model architectures.

5.3. Results

Performance comparisons among different architectures We tested various CNN backbones and compared the effect of adding one additional pathway or the additional task of predicting size information. To evaluate the model, we average root mean squared errors (RMSE), Pearson correlations, and accuracies for each action. We used Murray's method for calculating accuracies when evaluating the accuracy of regression models [22]. We first binarized the ground truth scores for each action and the predicted scores

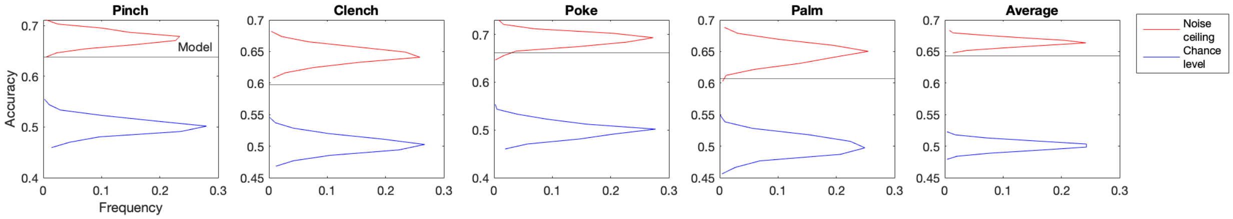
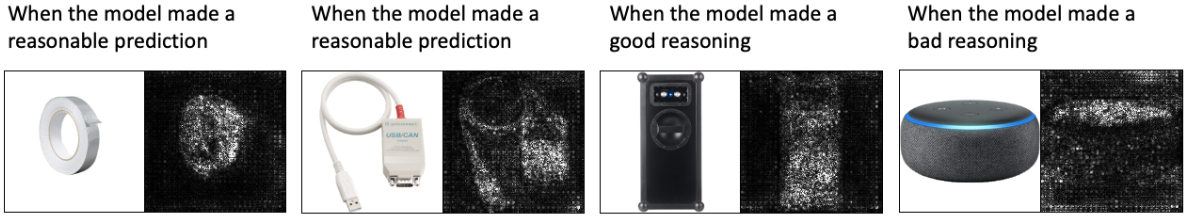


Figure 4. Theoretical upper (red) and lower (blue) bounds of the model performance for each of the affordance classes. The black lines show the best model (R2s in Table1) performance we achieved.


When the model made a reasonable prediction
When the model made a reasonable prediction
When the model made a good reasoning
When the model made a bad reasoning

Prediction	Pinch	Pinch	Palm	Clench
Ground truth	Clench	Clench	Clench	Poke

Figure 5. Example failure cases and the predictions' saliency maps. The failure cases could be categorized into three groups: (1) when the model made plausible prediction that differed from the label predictions (left two images), (2) when the model made atypical predictions, but with a plausible reason (second to right image), (3) and when the model made incorrect predictions without a good reason (the right image). The prediction and the ground truth actions in this plot correspond to the actions with the highest affordance scores from the model or from the annotations.

for each action using the ground truth scores medians as the thresholds. The accuracy was then calculated based on whether or not the ground truths and predictions were both higher (or lower) than the medians.

As shown in Table 1, the model using Resent152 as the backbone achieved the best performance out of all different types of CNN backbones [10]. Additionally, we can observe that by adding one additional pathway to preserve categorical information about the input object and requiring the network to solve an additional size prediction task, the model's performance improves. Furthermore, because our dataset contains multiple images of the same object (see Section 4 and Fig. 3b), we trained and tested our best-performing model on the subset of data that includes only one image for each object (the "no repeated objects" dataset). We found that our model performs well even this more challenging dataset (see Table 1).

Theoretical upper and lower bounds of the model performance Due to the fact that we are using newly created data, there is no existing baseline against which we can compare our proposed affordance score prediction algorithm. Therefore, we calculated theoretical upper and lower bounds that our algorithm could achieve on the dataset.

For the lower bound, we simulated a random agent 1000 times using a uniform distribution of hand action scores ranging from 0 to 100. Then, we calculated the accuracy of each hand action and average of those accuracies for each simulation. The blue lines in Fig. 4 are those simulated accuracies' distributions. As can be seen, the lower bounds are approximately 0.5.

We calculated the upper bound using the Monte Carlo noise ceiling (MCnc) method from functional magnetic resonance imaging literature [19]. To begin, we estimated the annotation measurement error (e.g., the motor control error in moving the sliding bar for annotations) by calculating variances of each annotator's familiarity scores for each image. This was possible because each Mturker annotated each image 4 times with the familiarity score. Then, we used the average of those variances as the estimated annotation measurement error. Following that, we used the normal distribution to estimate the hand action affordance distribution for each action in each image. We used sample means for each hand action score in each image as the center of the distribution and the difference between the sample variances for each hand action score and the estimated annotation measurement error as the distribution variance. Then, by sampling affordance scores from those estimated distributions, we simulated annotation by a single human annotator 1000 times and calculated the accuracies. The red distributions of Fig. 4 represent the accuracy distributions from the MCnc, the theoretical upper bound.

We can find from this upper- and lower-bound analysis

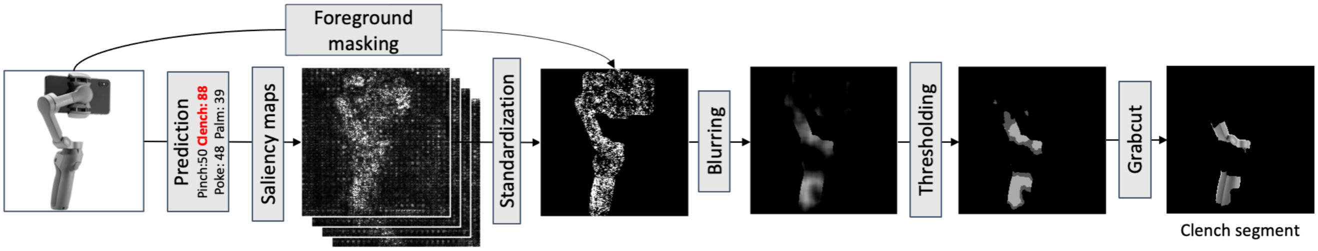


Figure 6. The segmentation pipeline. Here we see an example image going through the pipeline, where *clench* has the highest predicted action affordance score out of the four possible actions. Saliency maps of each action are generated, and segmentation output for *clench* is shown as an example. Note that we can also generate segmentation for other actions.

that our model performs significantly better than chance; it is almost comparable to the performance of a human.

Prediction performance by action type We can also observe from Fig. 4 that predicting the palm score is theoretically more unstable and difficult than predicting the other action scores. As previously stated in Section 4, the variance in palm score annotation was higher, or more inconsistent across annotators, than the variance in other actions' annotations. And, in contrast to the other action types, there was no image that had an average palm score that was significantly greater than the other scores. This is a natural consequence, given that majority of objects are intended to be used by pinching, clenching or poking.

On the other hand, the model performed the best at predicting poke scores. This, we presume, is because poke is frequently associated with buttons. As a result, it is assumed that the model learned to predict poke scores based on the presence of button-like structures.

Failure case analysis We first selected test images with the sum of the mean square errors for each action larger than 400 as *failure* cases. Then we could classify failure cases into 3 categories.

The first cases are when the model predicts affordance scores that differ from the labels but still make sense. The duct tape and the computer cable in Fig. 5 are examples of such cases. Humans can pinch those objects, and the model made predictions based on the locations that are actually pinch-able (brighter area).

The second cases are those in which the model makes peculiar predictions based on a legitimate reason. The speaker in the second to right image of Fig. 5 is an illustration of one of those. Although most people do not palm it, the model predicted a high palm score based on the speaker's flat surface area. As flat areas typically afford palm actions, this prediction was made based on a plausible reason [18].

The last cases are when the model made an incorrect prediction for no apparent reason. The mini speaker on the right of Fig. 5 is an example of this. While clenching the mini speaker is possible, the model predicted the high clench score based on the speaker's top, which can not be justified.

6. Weakly-Supervised Affordance Segmentation

6.1. Algorithm

We segment the objects to gets the parts that are most suitable for each hand action using the saliency maps produced using guided backpropagation [31]. In guided backpropagation, the gradients of hand action scores were calculated with respect to each input pixel, which was masked by the forward path activation. We used the absolute values of those gradients to construct saliency maps. As a saliency map demonstrates the network output's sensitivity to a unit change of each pixel in the original image, it locates the parts of the object that are the most important in deciding on particular hand actions. The saliency maps of each object was obtained when the object was in the test set during the 5-fold cross-validation.

After generating saliency maps for the four actions, we standardize each pixel of the saliency maps across the hand actions so that only the regions corresponding to each hand action remains. The foreground mask of the original image, which is based on maximum-area contour-finding, is multiplied by the standardized saliency map. We blur the output of the multiplication using a normalized box filter of size 20×20 so that the dots of the original image would be smoothed into larger contiguous patches. Then, we labeled pixels higher than 0.2 as the part of the candidate region, and pixels higher than 0.4 as part of the sure region of the object corresponding to the hand action. Finally, we apply GrabCut to the identified regions to obtain the segment of the image relevant to the hand action [28](Fig. 6).

6.2. Results

Fig. 7a displays successful examples of segmentation. For the first object (bell), the handle is segmented for *clench*. For the second object (calculator), the buttons are segmented for *poke*. Both of these segmentations are reasonable.

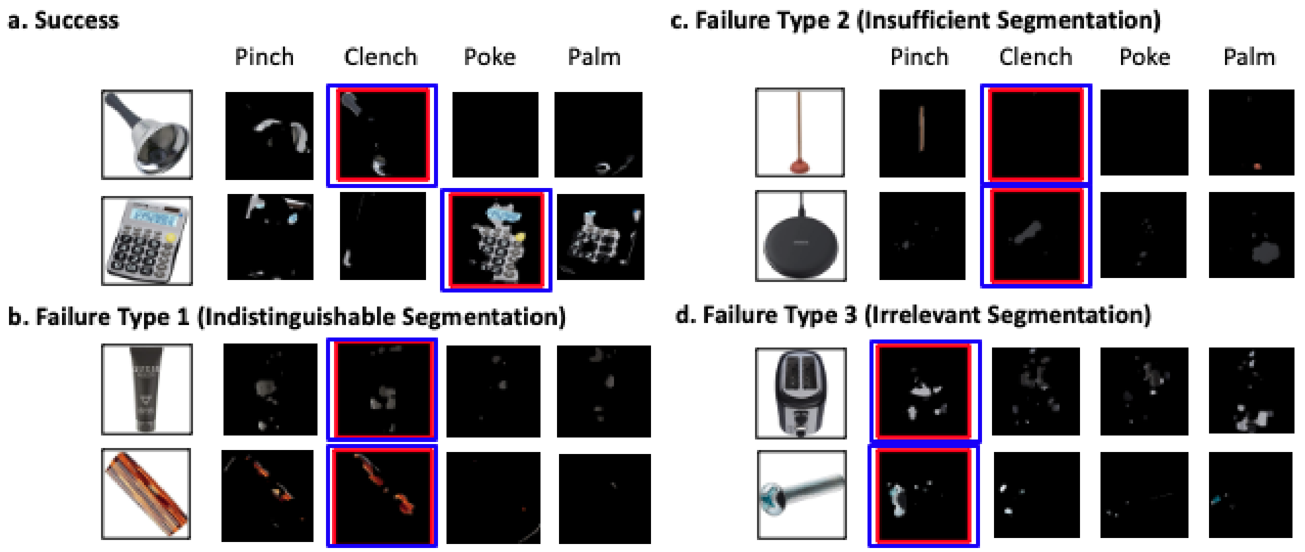

a. Success
Pinch
Clench
Poke
Palm
c. Failure Type 2 (Insufficient Segmentation)
Pinch
Clench
Poke
Palm
b. Failure Type 1 (Indistinguishable Segmentation)
d. Failure Type 3 (Irrelevant Segmentation)

Figure 7. The success and failure cases of segmentation. **(a)** shows successful cases. **(b)** shows failure of indistinguishable segmentation. **(c)** shows failure of insufficient segmentation. **(d)** shows failure of irrelevant segmentation. The red box is the hand action with the highest predicted affordance score, and the blue box is the hand action with the highest ground truth affordance score.

Failure cases are divided into three types. Fig. 7b demonstrates failure when the segmentation for pinch and clench are indistinguishable. One possible cause is the bias in the ground truth data for clench on objects that could be both pinched and clenched. Fig. 7c shows failure when segmented parts of objects are not large enough. The segmented portion of the first object (toilet plunger) is empty although the wood stick should have been segmented. The segmented portion of the second object (CD player) is small despite it can be held anywhere. Fig. 7d shows failure when segmentation seems irrelevant to the hand motion. For example, for the second object (screw), people usually pinch the body rather than the head. Yet, the head of the screw is segmented. Such a mistake may have been the result of the model using the shape of the object instead of its usage or other high-level information.

6.3. Human evaluation

Crowd-sourcing the annotations on the predicted segmentations To evaluate our segmentation model, we used human annotated ground truths collected from an Mturk task. Note that we did not collect the ground truth pixel-by-pixel segmentation but collected which is the most suitable action for the shown segmentations through multiple-choice questions where only one choice could be selected (For details, see Supplementary Fig. 1).

First, we randomly selected images from the Amazon dataset such that for each of these images, there was only one hand action out of the four hand actions that had a higher predicted score than the median of that hand action's ground truth scores. The predictions were made when the corresponding images were in the test sets during the 5-fold cross validation. We created stimuli out of a random 65-image subset (pinch=18, clench=20, poke=7, palm=20) of the Amazon images that passed this selection criterion.

For each trial, a stimulus is shown that is made up of two images, one placed above and one placed below (Fig. 8b, c, and d). The image above is the original image from the Amazon dataset, and the image below is the result of segmentation algorithm on the image that correspond to the action that was predicted by the prediction model with the highest predicted action score.

For each image that the Mturk worker is presented, they have to select the one hand action out of the four that they feel is the most suitable given what they see.

Three additional stimuli were presented during *catch trials* check Mturk workers' attention to the job (Supplemenatry Fig. 2). These were created by randomly taking three of the 65 stimuli and writing a message over the bottom image of the stimulus that prompted the reader to select a particular pre-specified hand action that was randomly selected from the four hand actions.

Next, we filtered out the annotation data from the workers that did not pass at least one of the catch stimulus trials and workers that failed any of the catch trials. Starting from the initial 61 Mturk workers who worked on our job, we filtered out 25 of the workers as bad performers. This resulted in 15.23 human annotations per image on average.

Crowd-sourced evaluation results Using only the data from the good workers, we calculated an accuracy metric

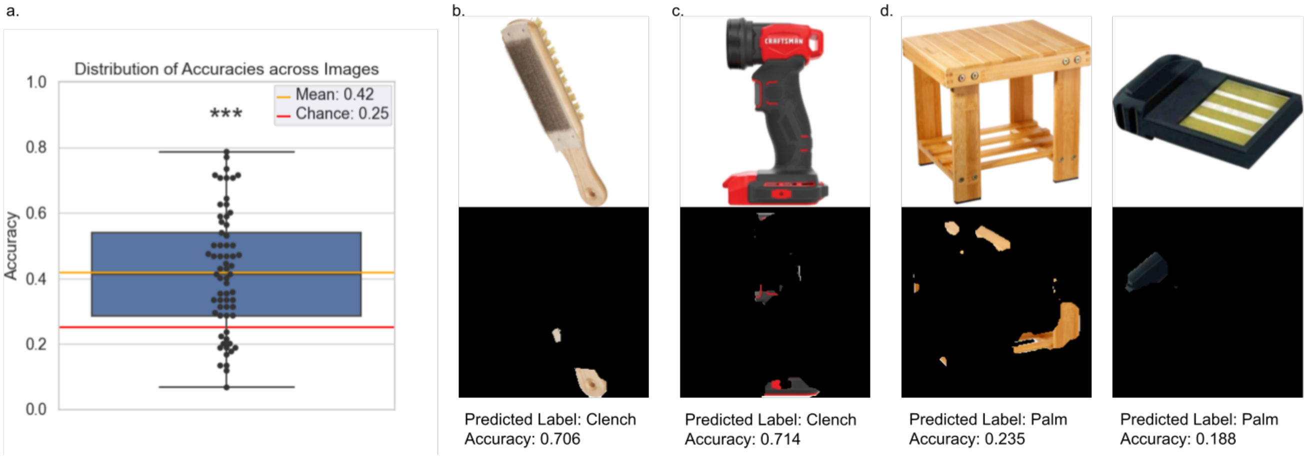


Figure 8. Human evaluation on the segmentation results. **(a)** The distribution of human annotation accuracy levels for each of the 65 stimuli, each dot representing one stimulus. **(b)** An example stimulus, which consists of the original image and the segmentation of the highest predicted action. Below the stimulus are the highest predicted action and the average human annotation accuracy level for that stimulus. **(c)** Here, the segmented regions seem to not be suited for the predicted label, but human annotations are still accurate at a level above 0.25. **(d)** Some example stimuli, predicted label, and accuracy levels, where the accuracy levels were below chance level. See Supplementary Fig. 3 for more examples and analyses.

for each of the 65 images. First, let "predicted label" be the action that was predicted by our prediction model to have the highest predicted score for a particular image. Second, let c_{match} be the total number of human annotations that agree with the predicted label for the image. Then, we calculated the image-specific accuracy as the ratio of c_{match} to the the total number of human annotations for the image.

The mean accuracy across all 65 images (0.42) was significantly higher than the chance-level accuracy (0.25) (one-sample t-test, $p < 10^{-9}, n = 68$)(See Fig. 8a). Some examples of stimuli, their predicted label, and the human annotation accuracy, for varying levels of accuracy, are shown in Fig. 8b, c, and d.

By repeating the same analysis separately on images that have the four different predicted labels we see that the human annotation accuracy levels are 0.37 ($p < 10^{-2}, n = 7$), 0.52 ($p < 10^{-6}, n = 20$), and 0.49 ($p < 10^{-5}, n = 18$), for the actions poke, clench, and pinch, respectively. These accuracy levels are all significantly above the chance level of 0.25. Meanwhile, for the group of images where the highest predicted action was palm, the average accuracy was 0.26 ($p = 0.35, n = 20$) (Supplementary Figs. 4 and 5).

However, we found that high human annotation accuracy of some images could be a confounded result. In images such as the ones shown in Fig. 8b and c, it is possible to guess the best hand action by looking only at the top image without looking at the segmentation below. For images where the original image displayed an object with a clear global hand action and where the bottom image showed the segments for the same hand action, we saw higher average accuracy.

7. Discussion

We proposed a CNN-based model that can predict hand action affordance of everyday objects as a many-to-many mapping. By utilizing a gradient-based CNN visualization method, the model could segment objects based on hand actions in a weakly-supervised manner without any additional training on the network.

To evaluate our algorithm quantitatively, we employed crowd-sourcing to assess our segmentation results. For a more robust evaluation on our segmentation method, we can also collect pixel-by-pixel annotations of hand action affordances on each image and calculate widely-used metrics in segmentation literature such as the intersection over union or the F1 score, which is planned for our future work.

The major limitation of our model is that it cannot generalize well to real-world images with cluttered backgrounds because the model was trained on images from Amazon with clean white backgrounds. Augmenting the white background with real world background images or adding images taken from wild to our dataset can be the solution to mitigate the problem.

As seen from our human evaluation results, the prediction and segmentation models are yielding results that match with human perception of affordance, at least for three of the four hand actions we focused on: pinch, clench, and poke. With further improvement, this hand action affordance prediction and segmentation model will provide a cost-effective method for detecting human hand action affordances, which can help propel the fields of robotic manipulation and ergonomic product design.

References

[1] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018.

[2] J. Bohg, A. Morales, T. Asfour, and D. Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2013.

[3] L. J. Buxbaum, K. M. Kyle, K. Tang, and J. A. Detre. Neural substrates of knowledge of hand postures for object grasping and functional object use: Evidence from fmri. *Brain research*, 1117(1):175–185, 2006.

[4] U. Castiello. The neuroscience of grasping. *Nature Reviews Neuroscience*, 6(9):726–736, 2005.

[5] M. M. de Wit, S. de Vries, J. van der Kamp, and R. Withagen. Affordances and neuroscience: Steps towards a successful marriage. *Neuroscience & Biobehavioral Reviews*, 80:622–629, 2017.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] T.-T. Do, A. Nguyen, and I. 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.

[8] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, volume 99, pages 1300–1309, 1999.

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

[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[11] T. Hermans, F. Li, J. M. Rehg, and A. F. Bobick. Learning contact locations for pushing and orienting unknown objects. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 435–442, 2013.

[12] P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[13] K. Iigaya, S. Yi, I. A. Wahle, K. Tanwisuth, and J. P. O'Doherty. Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features. *Nature Human Behaviour*, pages 1–13, 2021.

[14] L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, and J. Santos-Victor. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, 2016.

[15] Y. Jiang, S. Moseson, and A. Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *2011 IEEE International conference on robotics and automation*, pages 3304–3311. IEEE, 2011.

[16] S. H. Johnson-Frey. The neural bases of complex tool use in humans. *Trends in cognitive sciences*, 8(2):71–78, 2004.

[17] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.

[18] R. L. Klatzky, B. McCloskey, S. Doherty, J. Pellegrino, and T. Smith. Knowledge About Hand Shaping and Knowledge About Objects. *Journal of Motor Behavior*, 19(2):187–213, 1987-06.

[19] A. Lage-Castellanos, G. Valente, E. Formisano, and F. De Martino. Methods for computing the maximum performance of computational models of fmri responses. *PLoS computational biology*, 15(3):e1006397, 2019.

[20] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[21] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[22] N. Murray and A. Gordo. A deep architecture for unified aesthetic prediction. *arXiv preprint arXiv:1708.04890*, 2017.

[23] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[24] J. Napier, J. R. Napier, and R. H. Tuttle. *Hands*. Princeton University Press, 1993.

[25] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. 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.

[26] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. 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.

[27] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976.

[28] C. Rother, V. Kolmogorov, and A. Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[29] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

[30] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[32] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.