Ex Paucis Plura: Learning Affordance Segmentation from Very Few Examples

Johann Sawatzky, Martin Garbade, and Juergen Gall

University of Bonn
{sawatzky, garbade, gall}@iai.uni-bonn.de

Abstract. While annotating objects in images is already time-consuming, annotating finer details like object parts or affordances of objects is even more tedious. Given the fact that large datasets with object annotations already exist, we address the question whether we can leverage such information to train a convolutional neural network for segmenting affordances or object parts from very few examples with finer annotations. To achieve this, we use a semantic alignment network to transfer the annotations from the small set of annotated examples to a large set of images with only coarse annotations at object level. We then train a convolutional neural network weakly supervised on the small annotated training set and the additional images with transferred labels. We evaluate our approach on the IIT-AFF and Pascal Parts dataset where our approach outperforms other weakly supervised approaches.

1 Introduction

In order to use an object, an autonomous system has to precisely localize the parts of the object which are responsible for a certain type of interaction between this object and another object or the actor and the object. In comparison to object parts, affordances, which can be considered as functional attributes of objects, are more abstract since they generalize across object classes. Object parts from different object categories can share the same affordance class, if they share some similarity with respect to geometrical, categorical and physical properties, which in turn determines their functionality and usability. An example would be the shaft of a hammer and that of a tennis racket which would be both summarized under the class 'graspable'.

Recently several approaches have been proposed that use CNNs for detecting or segmenting affordances in images [18, 25, 34, 36, 26, 19]. However, the cost of data annotation constitutes a bottleneck for semantic segmentation in general and that of functional object parts in particular. While generating pixel-wise annotations of objects is already very time-consuming compared to bounding-box annotations, annotating even finer details like parts or affordances at large quantities as it is required for CNNs becomes infeasible.

In this work we show how to extend a tiny training set containing images with affordance annotations to make the training of a semantic segmentation CNN feasible. We assume that for the training set, we have only a handful of

images per object category (6 images per tool class in our experiments). For each image, the bounding box of the object and the the bounding boxes for all affordances of the object parts are given. Since training a CNN on such a small training set will be prone to overfitting, we make use of additional data where objects are already annotated by bounding boxes, but annotations of affordances or object parts are missing. We term the additional dataset as unlabeled since the images are not labeled in terms of affordances. Such data is already available, for instance, in form of object detection datasets.

In order to train the CNN on both datasets, *i.e.*, the small dataset with affordance annotations and the large dataset with only object annotations, we transfer the affordance annotations from the small dataset to the unlabeled images of the large dataset. For the label transfer, we use a semantic alignment network, which is trained without supervision, to find for each unlabeled image the most similar labeled image. Despite of having only bounding box annotations of affordances, we then train a CNN for pixel-wise affordance segmentation weakly supervised on both datasets. We evaluate our approach on the IIT-AFF dataset [26] and the Pascal Parts dataset [6] where our approach outperforms other segmentation approaches that are also trained weakly supervised.

2 Related Work

Most related to our work are approaches for weakly supervised semantic segmentation and approaches for affordance detection or segmentation.

The task of weakly supervised semantic segmentation is to learn pixel-wise classification from a more coarse level of supervision. The different approaches vary in the type of supervision cues: [29, 30, 32] use image level labels as supervision to train semantic segmentation models while casting weakly supervised learning as a constrained optimization or a multiple instance learning problem, respectively. [20, 2] leverage user annotated scribbles and individual key points to provide either sparse object labels or object location priors. [28, 12] apply expectation maximization for weakly supervised training. While both approaches use image level labels, [28, 13, 7] additionally use annotated object bounding boxes while [12, 27] uses saliency masks for supervision. Another paradigm, called simple-to-complex [41, 38], consists of first training a model using simple images, *i.e.*, images containing a single object category followed by the training on complex images, *i.e.*, images with multiple objects. [41] combine an object proposal generator [1] with a proposal selection module thereby linking semantic segmentation and object localization. [16] improves the training procedure by using multiple loss functions. [3] combines saliency and attention maps to approximate the ground truth annotation. Some approaches explore the concept of region based mining, *i.e.*, an initial localization seed is expanded to the size of objects [14, 40]. While these works address object segmentation, a few works focus on weakly supervised semantic part detection [17] or segmentation [23].

Various image domains have been explored for affordance detection or segmentation. The context of affordances also strongly differs depending on the

task such as understanding human body parts [21], classifying environment affordances [34, 31], or detecting affordances of real world objects that robots interact with [24]. [37] use predefined primary tools to infer object functionalities from 3D point clouds. [39] detect grasp affordances by combining the global object poses with its local appearance. [15] detect object affordances by observing object-action interactions performed by humans. Recently there have been several works that rely on deep convolutional neural networks for affordance detection [18, 25, 34, 36, 26, 19]. [11] propose a region alignment layer to align the input image space with the feature map space. [8] detect multiple affordance classes in the object, instead of binary classes as in [11].

3 Label Transfer for Affordance Segmentation

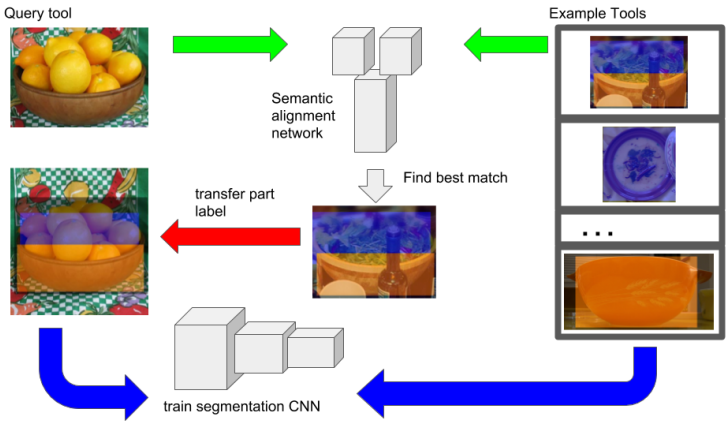

Fig. 1. In order to train a network to segment affordances from a very small set of examples, we transfer labels to unlabeled images. The training data consists of a set of objects where affordances are annotated by bounding boxes (right). This training set is very small and comprises only a few examples per object category. We then collect more examples of objects from an object detection dataset, i.e., the bounding box and the name of the object are given but not the affordances (left). To transfer the annotation labels of the training set to the new images, we use a semantic alignment network to find for each new image the most similar image in the training set. The bounding box annotations of the affordances are then transferred to the matched images and a CNN is then trained on all images. Best viewed in color.

Since annotating affordances or parts of objects is time-consuming, our goal is to train a convolutional network that segments affordances in images on a very small set of annotated images and additional unlabeled images. An overview of the approach is given in Fig. 1.

Our training set consists of a few example images for each object category where the affordances are annotated by bounding boxes. Since large datasets for object detection exist, we make use of them to extend the training set. These datasets, however, do not provide any annotations of affordances or parts but only bounding boxes for the objects. We therefore transfer the affordance labels from our training set to the objects from an object detection dataset. To this end, we first use a semantic alignment network to retrieve for each unlabeled image the most similar annotated training image to transfer the annotations (Section 3.1). We then train a fully convolutional network on the original training set and the extended set with transferred labels and use this model for inference (Section 3.2).

3.1 Semantic Alignment Network for Similarity Estimation

For similarity matching between annotated example objects and unannotated query objects, we use the semantic alignment network proposed in [33]. It takes two images I_s and I_t and predicts an affine transformation T_{aff} and a thin plate spline transformation T_{tps} whose concatenation semanticly aligns I_s to I_t. The transformations are subsequently predicted by two networks only differing in the final layer.

First the feature maps of both images $f^s_{ij:}$ and $f^t_{kl:}$, where i and j are the spatial coordinates in the source image I_s and k and l are the spatial coordinates in the target image I_t, are extracted in two Siamese branches. Then, a 4D-tensor S of space match scores is obtained via

$$s_{ijkl} = \frac{\langle f^s_{ij:}, f^t_{kl:} \rangle}{\sqrt{\sum_{a,b} \langle f^s_{ab:}, f^t_{kl:} \rangle^2}}. \tag{1}$$

Next, the parameters G of the geometrical transformation T_G are calculated from the space match score tensor S. This yields then the 4D inlier mask tensor M:

$$m_{ijkl} = \begin{cases} 1 & \text{if } d(T_G(i,j),(k,l)) < \tau \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where d is the Euclidean distance. For τ, we use the same value as in [33]. Combining M and S provides the soft inlier count, a measure for the quality of the alignment:

$$c = \sum_{i,j,k,l} s_{ijkl} m_{ijkl}. \tag{3}$$

Intuitively, the feature vectors of pixels in the target and the warped image should be similar if the points are spatially close ($m_{ijkl} = 1$). Therefore, $-c$ serves as a training loss.

We use the pre-trained model [33], which has been first trained on synthetic data obtained from the Pascal dataset [9] and then finetuned with image pairs

from the PF-PASCAL dataset [10]. Since the loss does not require human supervision and the network does not explicitly take any note of the object class, the model generalizes to unseen object classes. We therefore can use the model trained on Pascal classes on the IIT-AFF dataset.

The approach, however, fails for large transformations. Already 2D rotations by more than 30 degrees lead to poor semantic alignments. We therefore augment each of the annotated examples by rotating it by 90, 180 and 270 degrees and flipping it. To find the best match in our annotated training set $\{I_i\}_{i\in\{1,\dots n\}}$for a query image J, we compute (3) for J and each image I_i, which contains the same object class as J. The best match for J is then given by the image I_k with the highest soft inlier count c.

Fig. 2 shows some examples for the top 3 matches.

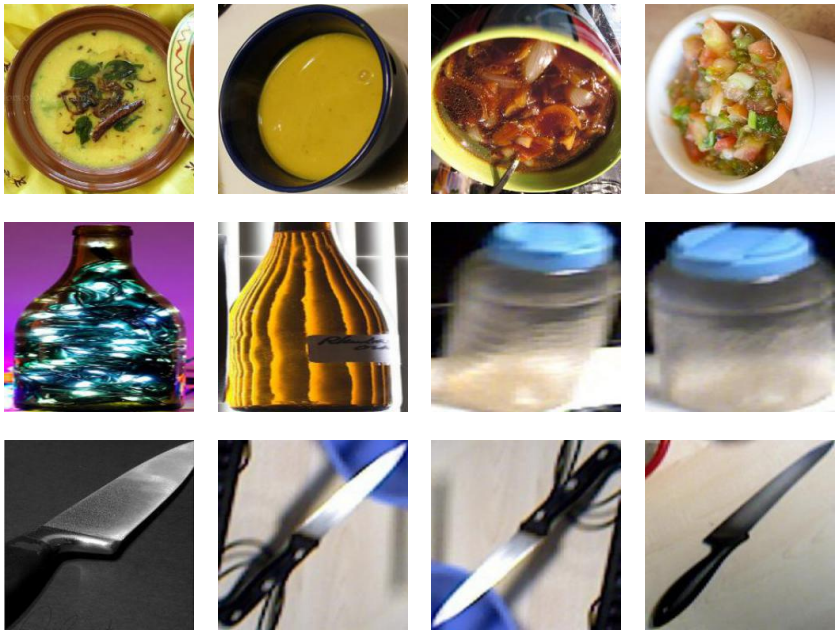

Fig. 2. Some query tools (left column) and the top 3 matching example tools with decreasing proximity from left to right. Except for the second match for the knife, the matching procedure retrieves tools seen from a similar viewpoint and having same orientation. Best viewed in color.

In order to transfer the affordance labels from I_k to J, the estimated warping transformation could be used. However, our experiments reveal that the estimated transformations are not accurate enough for transferring the labels. Instead, we scale I_k to match the size of J and copy the annotations from I_k to J.

3.2 Semantic Segmentation

In our experiments, we will investigate two supervision levels and two transfer strategies. In the first supervision setting, the affordances of example tools are pixel-wise annotated. In the second setting, the affordances of example tools are annotated by bounding boxes. In the latter case, we obtain a rough pixel-wise annotation by setting all pixel labels inside an annotated bounding box to its affordance class. If a pixel is located inside multiple affordance bounding boxes, it receives the affordance label of the smallest bounding box. We refer to these supervision levels by *bbox* and *pixelwise*. The *copy* strategy simple resizes and copies the labels of the example tool onto the query tool. The *warp* strategy warps the label of the example tool using the transformation predicted by the alignment network. For both transfer strategies, all pixels located outside the object bounding boxes are set to background and all pixel labels inside the object bounding boxes which were not assigned to an affordance class are ignored and thus do not contribute to the loss when training the semantic segmentation network. We combine the notations of supervision level and transfer strategy, for example *bbox-copy* means *bbox* supervision level and *copy* transfer method. Fig. 3 illustrates our supervision levels and transfer strategies. The proposed method assumes *bbox* for supervision and uses *copy* for label transfer. For semantic segmentation, we use the deeplab VGG architecture [4], which is a fully convolutional network providing as output a feature map f with width and height equal to the input image and each channel corresponding to an affordance or background. We obtain the affordance probability by taking the pixelwise softmax of f. During training, the loss for a particular pixel is computed individually. If the ground truth label is an affordance or background it equals the cross entropy between the ground truth label and the prediction, otherwise it is 0. The overall loss is the sum of the pixel-wise losses. During inference, we use the conditional random field layer of deeplab on top of the final feature map.

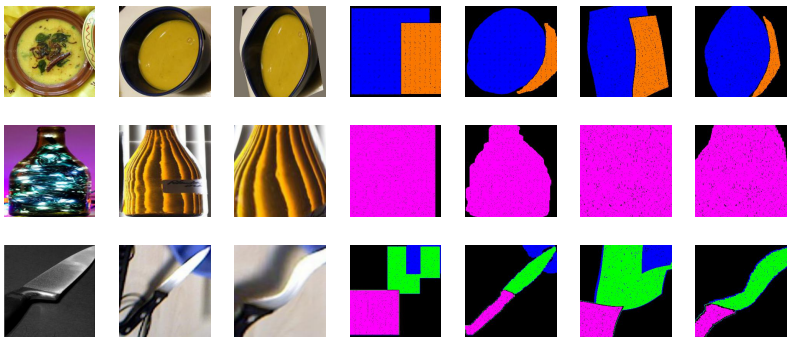

Fig. 3. Illustration of our supervision levels and transfer strategies. From left to right: Query tool, matched example tool, aligned example tool, *bbox-copy* labels, *pixelwise-copy* labels, *bbox-warp* labels, *pixelwise-warp* labels. Best viewed in color.

4 Experiments

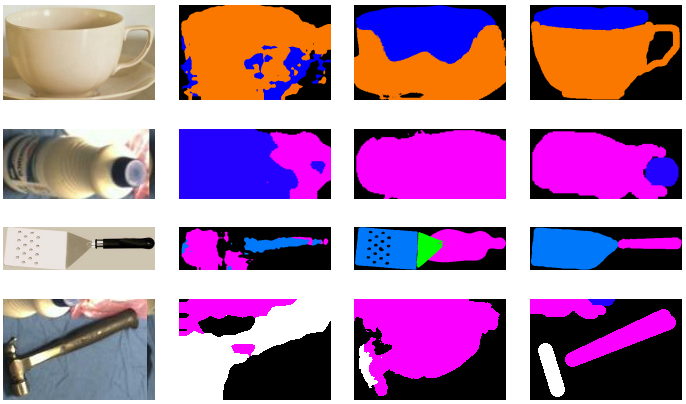

Fig. 4. Qualitative results on bounding boxes: RGB input (first column), DCSP [3] results (second column), our results (third column), ground truth (last column). In contrast to DCSP, our method correctly associates the affordances with the respective object parts. Best viewed in color.

We conduct our experiments on the IIT-AFF dataset introduced by [26]. It consists of images showing 10 classes of tools in context. There are 6184 images in the trainval set and 2651 images in the test set. The images were collected in a robotics lab or come from the Imagenet dataset [35]. Each tool is annotated with a bounding box. Additionally, each tool class has a predefined set of possible affordances. Tool parts serving an affordance are pixel-wise annotated with it. The tool classes with their affordances are: bowl (wrap, contain), tv (display), pan (contain, grasp), hammer (grasp, pound), knife (cut, grasp), cup (contain, wrap), drill (grasp, engine), racket (grasp, hit), spatula (support, grasp) and bottle (grasp, contain).

Unless stated otherwise, in all our experiments our unlabeled set comprises the images from the IIT trainval set with 6 example tools per tool class randomly drawn from them to constitute the training set. We use the semantic alignment model trained on PF-PASCAL [10] by [33]. For training and inference with deeplab [4], we use the same setup as in the original paper in the fully supervised setup on Pascal.

4.1 Comparison to State of the Art

To our knowledge there is no work on weakly supervised semantic segmentation which uses the same amount of supervision: A vast object dataset annotated on bounding box level but unlabeled in terms of object part affordances and a tiny dataset with bounding boxes provided for affordances. DCSP [3], which is

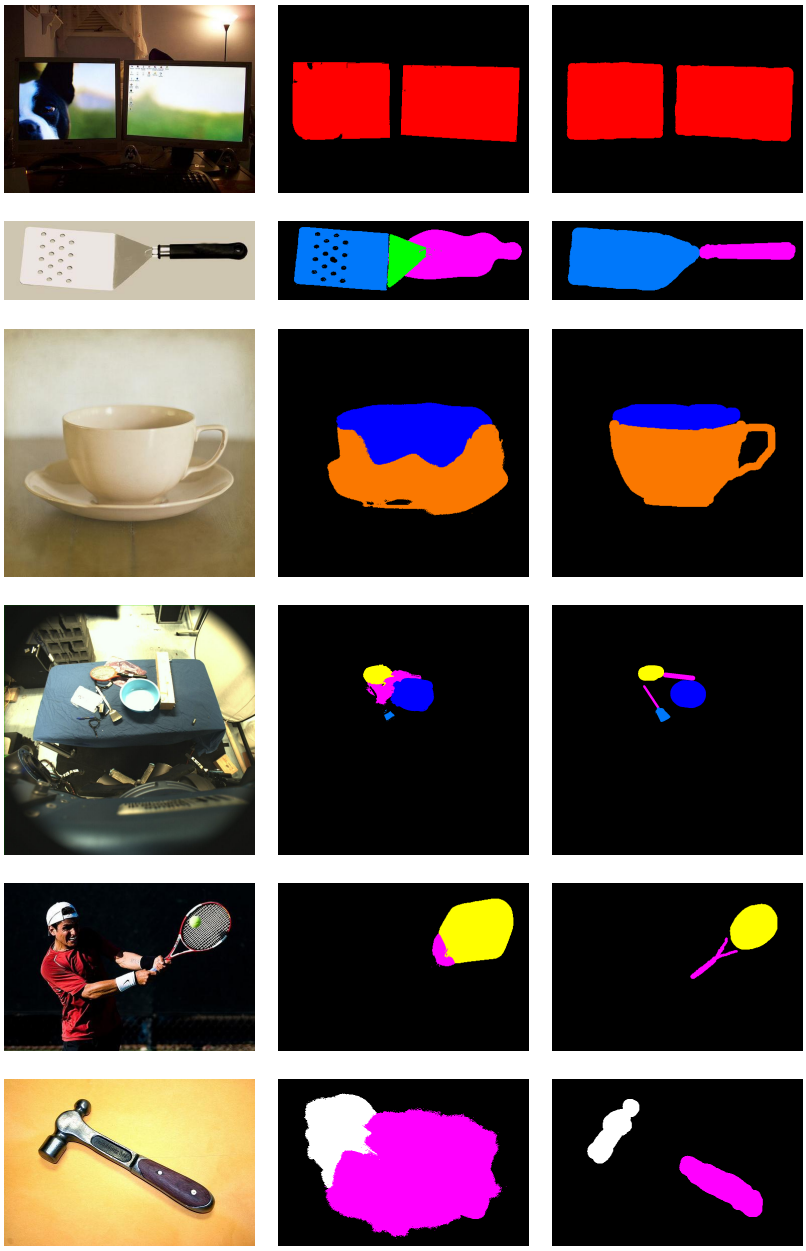

Fig. 5. Qualitative results on IIT-AFF [26]: RGB input (left), our results (middle), ground truth (right). Best viewed in color.

the current state of the art method for weakly supervised image segmentation on Pascal VOC 2012, uses a list of present classes in an image for supervision. We therefore train DCSP on the bounding boxes of tools from the unlabeled set as well as on the annotated bounding boxes of affordances from the training set. On the unlabeled set, the affordances are inferred from the tool class and used as image labels for supervision. We keep the original training parameters of DCSP, but reduce the learning rate by a factor of 3 since it improved the results of DCSP.

For comparison, we evaluate both methods not on the entire images, but only within the annotated bounding boxes surrounding the objects, since we are interested in how well both methods segment the affordances within a bounding box. The results are reported in Table 1: DCSP achieves a mean IoU of 29.7% while our approach yields 46.1%, thus outperforming DCSP by more then 16%. To analyze if the difficulty for DCSP stems from the localization of affordances on the tool or from the pixel-wise segmentation of the tool itself, we performed an additional experiment. Instead of training DCSP for segmenting affordances, we trained DCSP for segmenting objects. For object segmentation, DCSP performs very well and achieves 53.4% mean IoU for the object categories. Therefore localizing the affordance on the tool constitutes the main challenge. This is also evident from the qualitative comparison shown in Fig. 4. Qualitative results for complete images are shown in Fig. 5.

Table 1. Comparison to DCSP [3], a method showing state of the art results on Pascal VOC 2012. We report IoU on the IIT-AFF dataset [26]. For a fair comparison, we train and evaluate DCSP on bounding box crops of tools and the affordance segments of example tools and evaluate ours on the bounding box crops only.

method	contain	cut	display	engine	grasp	hit	pound	support	wrap grasp	mean
DCSP [3]	0.340	0.179	0.602	0.214	0.259	0.548	0.242	0.085	0.205	0.297
proposed	0.616	0.209	0.811	0.364	0.328	0.633	0.345	0.335	0.510	0.461

4.2 Number of Examples

Our proposed evaluation setup uses 6 random examples per tool class, but we also investigated the performance with an even smaller amount of examples, namely 1, 2 and 3 examples per tool class, and report the results in Table 2. Unlike in the previous section, we evaluate on complete test images, but still achieve a mean IoU of 41.9%. When using only one example tool per tool class, the performance drops to 35.4%.

4.3 Impact of Additional Training Data

To see if additional training data and transferring the labels from the examples to the additional training data is required at all, we trained our semantic

Table 2. Evaluation of our method on the IIT-AFF dataset [26] for different number of example tools per tool class. We evaluate on full images and report IoU.

# examples	contain	cut	display	engine	grasp	hit	pound	support	wrap-gr.	mean
1	0.480	0.152	0.760	0.305	0.293	0.575	0.101	0.134	0.387	0.354
2	0.518	0.191	0.744	0.336	0.267	0.590	0.248	0.096	0.426	0.380
3	0.522	0.182	0.716	0.331	0.269	0.645	0.245	0.136	0.417	0.385
6 (default)	0.564	0.180	0.723	0.329	0.288	0.596	0.295	0.323	0.469	0.419

segmentation network only on the images containing at least one of the 60 example tools. Pixels located inside the affordance bounding box of the example tools were set to this affordance class, pixels belonging to any tool bounding box which does not belong to an example tool were ignored during training, and the rest was set to background. Since the number of training images is tiny in this setting, we reduced the number of iterations from 6000 to 300 and the step length during training accordingly to avoid overfitting. As can be seen in Table 3, the performance drops to 27.3% and for the affordances cut, pound, support to almost 0. Our approach is especially beneficial for challenging small affordances.

Table 3. Comparison of training on the example images only vs. our approach, which uses additional training data by label transfer. We evaluate on full images from the IIT-AFF dataset [26] and report IoU.

method	contain	cut	display	engine	grasp	hit	pound	support	wrap-gr.	mean
ex. tools only	0.563	0.000	0.501	0.206	0.226	0.553	0.005	0.016	0.388	0.273
proposed	0.564	0.180	0.723	0.329	0.288	0.596	0.295	0.323	0.469	0.419

4.4 Warping vs No Warping

While we simply resize and copy the affordance localization cues from the most similar example tool to the tool of interest, one could also warp the localization cues of the example tool onto the target tool using the transformation provided by the semantic alignment network. On the one hand, this approach has the advantage of potentially better aligning the shape of the tools and therefore better aligning the functional parts. On the other hand, the warping might be reasonable for only some parts of the object but fail for other, in particular small parts. Therefore, the benefit of using the warping transformation or not depends on the affordance classes. The results reported in Table 4 show that warping improves the accuracy for the classes display, engine, and hit, but it decreases the accuracy for the other affordance classes. In average, using the estimated warping transformation for label transfer reduces the accuracy from 41.9% to 39.1%.

Table 4. Comparison of warping the affordance labels from the example tool vs copying them. We evaluate on full images from the IIT-AFF dataset [26] and report IoU.

method	contain	cut	display	engine	grasp	hit	pound	support	wrap-gr.	mean
bbox-warped	0.550	0.156	0.730	0.313	0.278	0.640	0.188	0.218	0.443	0.391
proposed	0.564	0.180	0.723	0.329	0.288	0.596	0.295	0.323	0.469	0.419

4.5 Bounding Box vs. Pixel-wise Annotation

Obtaining affordance region bounding boxes for example tools is far cheaper than annotating the functional regions pixel-wise. To investigate the potential gain from a pixel-wise annotation, we conducted two ablation experiments. In the first, we transfer the pixel-wise affordance annotations from example tools to unlabeled tools without using the estimated warping transformation and in the second we use the estimated warping transformation for label transfer. We report the results in Table 5: Providing pixel-wise affordance annotations for example tools increases the accuracy with and without warping. In case of warping the accuracy increases from 39.1% to 40.1% and without warping the accuracy increases from 41.9% to 44.8%.

Table 5. Comparison of using accurate pixel-wise affordance annotations of the example tools vs. bounding boxes around affordances. We report the results with and without using the estimated warping transformation for label transfer. We evaluate on full images from the IIT-AFF dataset [26] and report IoU.

method	contain	cut	display	engine	grasp	hit	pound	support	wrap-gr.	mean
pxlwise copy	0.601	0.245	0.745	0.368	0.388	0.589	0.260	0.333	0.502	0.448
pxlwise warp	0.592	0.190	0.748	0.363	0.354	0.616	0.0	0.278	0.466	0.401
proposed	0.564	0.180	0.723	0.329	0.288	0.596	0.295	0.323	0.469	0.419

4.6 ResNet Features vs. Alignment

To investigate the benefit of the unsupervisedly trained semantic alignment network, we train a semantic segmentation model using an approach identical to the proposed method except for the matching criterion between query tools and example tools. Since the alignment network was trained on Pascal VOC2012, we take the Pascal VOC2012 semantic segmentation Resnet-101 model from [5], and generate the features of the res5c layer for each query tool and each example tool. Note that we use the same CNN backbone as for the semantic alignment network and require the same amount of cross dataset generalisation. After that, we retrieve for each query tool the example tool with the most similar feature map and transfer the labels. Specifically, the cosine distance serves as a measure

for the similarity of the vectorized feature maps of two images v, w:

$$d = 1 - \frac{\langle v, w \rangle}{\|v\|\|w\|} \tag{4}$$

As can be seen from Table 6, the ResNet-101 features perform slightly worse than the weak alignment network.

Table 6. Comparison of two matching strategies between query tools and example tools: The proposed strategy uses the loss of a semantic alignment network trained in an unsupervised manner, the ablation uses the features of ResNet-101 pretrained on Pascal VOC2012. We evaluate on full images from the IIT-AFF dataset [26] and report IoU.

	contain	cut	display	engine	grasp	hit	pound	support	wrap grasp	mean
features	0.573	0.206	0.705	0.348	0.287	0.608	0.262	0.322	0.423	0.415
proposed	0.564	0.180	0.723	0.329	0.288	0.596	0.295	0.323	0.469	0.419

4.7 Oracle Experiment: Ground Truth Bounding Box for Each Affordance of Each Query Tool

In this ablation experiment we investigate what is achievable if the bounding boxes around affordances are not only given for the example tools but also for all query tools. All pixels inside an affordance bounding box are set to this affordance. In case of a pixel belonging to multiple affordance bounding box, it is assigned to the affordance with the smallest bounding box. All other pixels are set to background. After that, the semantic segmentation network is trained and used for inference as in our proposed method. We report the results in Table 7. This additional supervision improves the results to 52.6%, however, at the cost of additional annotations of query tools, while our method does not require any additional annotation once object bounding boxes are given. For example, it could be applied to the affordances of objects in the COCO dataset [22]. Even if the bounding boxes are not given for a custom data set, they can be generated using a weakly supervised object detection system, e.g. [42].

Table 7. Results if ground truth bounding boxes would be given for each affordance of each query tool vs. our method. We evaluate on full images from the IIT-AFF dataset [26] and report IoU.

	contain	cut	display	engine	grasp	hit	pound	support	wrap grasp	mean
gt-bbox	0.686	0.217	0.747	0.521	0.389	0.722	0.382	0.466	0.606	0.526
proposed	0.564	0.180	0.723	0.329	0.288	0.596	0.295	0.323	0.469	0.419

4.8 Evaluation on the Pascal Parts dataset

We finally evaluate our approach on the Pascal Parts dataset [6]. It contains images from the Pascal VOC dataset, which belong to the categories bird, cat, cow, dog, horse, person, and sheep. For each category, 4 to 5 semantic body parts are annotated. The task of part segmentation differs from affordance segmentation since different object classes do not share the same part category, *e.g.*, leg of horse and leg of sheep are considered as two different part classes in Pascal Parts. This is in contrast to affordances, which are shared among different tool classes. Since our method can be applied to both tasks, we also evaluate our approach on this dataset by randomly sampling 6 example objects per object class. Our approach outperforms the current state of the art by +3.5% as can be seen in Table 8.

Table 8. Evaluation on the Pascal Parts [6]. Our method outperforms state of the art methods for weakly supervised semantic parts segmentation. As on IIT-AFF dataset [26], we use 6 example objects per object class.

method	Bird	Cat	Cow	Dog	Horse	Person	Sheep	mean
[17]	0.099	0.135	0.115	0.141	0.067	0.106	0.105	0.110
[23]	0.111	0.113	0.124	0.142	0.075	0.128	0.106	0.114
proposed	0.148	0.174	0.115	0.180	0.120	0.108	0.201	0.149

5 Conclusion

In this work, we have shown that a CNN, which is weakly supervised trained for affordance or object part segmentation, can be trained from very few annotated examples. This has been achieved by exploiting a semantic alignment network to transfer annotations from a small set of annotated examples to images that are only annotated by the object class. In our experiments, we have shown that our approach achieves state of the art accuracy on the IIT-AFF dataset [26] and the Pascal Parts dataset [6].

Acknowledgement

The work has been financially supported by the DFG projects GA 1927/5-1 (DFG Research Unit FOR 2535 Anticipating Human Behavior).

References

1. Arbeláez, P.A., Pont-Tuset, J., Barron, J.T., Marqués, F., Malik, J.: Multiscale combinatorial grouping. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 328–335. IEEE Computer Society (2014)

2. Bearman, A.L., Russakovsky, O., Ferrari, V., Li, F.: What's the point: Semantic segmentation with point supervision. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII. pp. 549–565. Springer (2016)
3. Chaudhry, A., Dokania, P.K., Torr, P.H.S.: Discovering class-specific pixels for weakly-supervised semantic segmentation. In: British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017. BMVA Press (2017)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. International Conference on Learning Representations (2015)
5. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR **abs/1606.00915** (2016)
6. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.L.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 1979–1986. IEEE Computer Society (2014)
7. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1635–1643. IEEE Computer Society (2015)
8. Do, T., Nguyen, A., Reid, I.D., Caldwell, D.G., Tsagarakis, N.G.: Affordancenet: An end-to-end deep learning approach for object affordance detection. CoRR **abs/1709.07326** (2017)
9. Everingham, M., Eslami, S.M.A., Van Gool, L.J., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015)
10. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow: Semantic correspondences from object proposals. IEEE Trans. Pattern Anal. Mach. Intell. **40**(7), 1711–1725 (2018)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2980–2988. IEEE Computer Society (2017)
12. Hou, Q., Dokania, P.K., Massiceti, D., Wei, Y., Cheng, M., Torr, P.H.S.: Mining pixels: Weakly supervised semantic segmentation using image labels. CoRR **abs/1612.02101** (2016)
13. Khoreva, A., Benenson, R., Hosang, J.H., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1665–1674. IEEE Computer Society (2017)
14. Kim, D., Cho, D., Yoo, D.: Two-phase learning for weakly supervised object localization. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 3554–3563. IEEE Computer Society (2017)
15. Kjellström, H., Romero, J., Kragic, D.: Visual object-action recognition: Inferring object affordances from human demonstration. Computer Vision and Image Understanding **115**(1), 81–90 (2011)
16. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling,

M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV. pp. 695–711. Springer (2016)
17. Krause, J., Jin, H., Yang, J., Li, F.: Fine-grained recognition without part annotations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 5546–5555. IEEE Computer Society (2015)
18. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. I. J. Robotics Res. **34**(4-5), 705–724 (2015)
19. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4438–4446. IEEE Computer Society (2017)
20. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 3159–3167. IEEE Computer Society (2016)
21. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5168–5177. IEEE Computer Society (2017)
22. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. pp. 740–755. Springer (2014)
23. Meng, F., Li, H., Wu, Q., Luo, B., Ngan, K.N.: Weakly supervised part proposal segmentation from multiple images. IEEE Trans. Image Processing **26**(8), 4019–4031 (2017)
24. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015. pp. 1374–1381. IEEE (2015)
25. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Detecting object affordances with convolutional neural networks. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016. pp. 2765–2770. IEEE (2016)
26. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: 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 2017, Vancouver, BC, Canada, September 24-28, 2017. pp. 5908–5915. IEEE (2017)
27. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5038–5047. IEEE Computer Society (2017)
28. Papandreou, G., Chen, L., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1742–1750. IEEE Computer Society (2015)

29. Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1796–1804. IEEE Computer Society (2015)
30. Pathak, D., Shelhamer, E., Long, J., Darrell, T.: Fully Convolutional Multi-Class Multiple Instance Learning. International Conference on Learning Representations Workshop (2015)
31. Pham, T., Do, T.T., Sünderhauf, N., Reid, I.: SceneCut: Joint Geometric and Object Segmentation for Indoor Scenes (2018)
32. Pinheiro, P.H.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1713–1721. IEEE Computer Society (2015)
33. Rocco, I., Arandjelović, R., Sivic, J.: End-to-end weakly-supervised semantic alignment. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, USA, June 19 -21, 2018 (2018)
34. Roy, A., Todorovic, S.: A multi-scale CNN for affordance segmentation in RGB images. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV. pp. 186–201. Springer (2016)
35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
36. Sawatzky, J., Srikantha, A., Gall, J.: Weakly supervised affordance detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5197–5206. IEEE Computer Society (2017)
37. Schoeler, M., Wörgötter, F.: Bootstrapping the semantics of tools: Affordance analysis of real world objects on a per-part basis. IEEE Trans. Cognitive and Developmental Systems **8**(2), 84–98 (2016)
38. Shen, T., Lin, G., Liu, L., Shen, C., Reid, I.D.: Weakly supervised semantic segmentation based on co-segmentation. In: British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017. BMVA Press (2017)
39. Song, H.O., Fritz, M., Göhring, D., Darrell, T.: Learning to detect visual grasp affordance. IEEE Trans. Automation Science and Engineering **13**(2), 798–809 (2016)
40. Wei, Y., Feng, J., Liang, X., Cheng, M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6488–6496. IEEE Computer Society (2017)
41. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence (2017)
42. Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B.: W2f: A weakly-supervised to fully-supervised framework for object detection. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, USA, June 19 -21, 2018. IEEE Computer Society (2018)