One-Shot Transfer of Affordance Regions? AffCorrs!

Anonymous Author(s)
Affiliation
Address
email

Abstract: In this work, we tackle one-shot visual search of object parts. Given a single reference image of an object with annotated affordance regions, we segment semantically corresponding parts within a target scene. We propose AffCorrs, an unsupervised model that combines the properties of pre-trained DINO-ViT's image descriptors and cyclic correspondences. We use AffCorrs to find corresponding affordances both for intra- and inter-class one-shot part segmentation. This task is more difficult than supervised alternatives but enables future work such as learning affordances via imitation and assisted teleoperation.

Keywords: One-Shot, Affordance, Correspondence

1 Introduction

Robot agents can significantly benefit from perceiving and understanding what the environment *affords* them to do. Affordances [1] are representations of how a part of the environment can be used, e.g., a spoon *affords* to be grasped, and to contain liquid. Being grounded on human semantics, they are intuitive and explainable. Part-based affordance representations can be efficiently used as an intermediate representation that reduces the dimensionality of many robot learning problems significantly [2]. Just as people are able to transfer the knowledge of an object's functionality to other objects from only a few examples, e.g., the graspable handles of jugs to mugs, it would be beneficial for robots to understand such correspondences, too. We motivate the topic of recognizing known affordance regions in unseen novel objects as a useful step toward more autonomous robots, assisted teleoperation, visual inspection, and scene understanding.

Evidently, semantic part correspondence can be achieved with fully supervised methods [3, 4]. However, they are limited to objects similar to the ones present in the dataset, each requiring multiple instances for better generalization of the object category. Meanwhile, the self-supervision and unsupervised learning paradigms, present an alternative direction, which alleviates the data annotation bottleneck [5], and could enable robots to learn continually by themselves [6].

In this work, we demonstrate how the pre-trained DINO-ViT model [7], which is shown to produce good co-segmentation and point correspondences [8], can be used for part querying and finding semantically corresponding parts in one-shot – a novel formulation of the One-Shot Instance Segmentation (OSIS) problem extended to parts, which we call One-Shot Affordance Part Segmentation (OSAPS). The ability to query on particular parts leverages the semantic prior given by the user, or alternatively by a preceding system. We demonstrate how these parts can be associated with predefined skills. As far as we are aware, this work is one of the first to extend one-shot instance segmentation to part masks, and affordance parts in particular. Our proposed method presents several contributions: (i) **One-Shot** affordance transfer instead of supervised learning; (ii) **Decoupled representation** that allows the stack of affordance skills to be separate from the affordance discovery; (iii) **Query-based** transfer instead of co-part segmentation, which allows us to transfer specific semantic parts rather than parts based on visual or geometric features only; (iv) **Benchmark** subset of an affordance dataset curated for one-shot part transfer.

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

2 Related Work

Most affordance learning work focuses on fully supervised methods that learn a particular affordance, predefined in the training set as image segments, contact points or interaction regions [9, 10, 11]. Some works focus on learning from interaction combined with geometric and perceptual features [12, 13]. Model-based approaches are a potential path towards efficient knowledge transfer. However, they still rely on expensive annotations, such as full object template models [14]. One-shot affordance detection was recently demonstrated [15], but this approach is limited to instance masks and bounding boxes, which are insufficient to transfer knowledge to a robot.

Models requiring little to no supervision have great potential for robot applications that benefit from adapting quickly and learning continuously. It is of particular interest to understand how previous works in computer and robot vision solve the semantic correspondence problem in the domains of point- and part-correspondence with limited supervision.

Many works enforce the semantic relationship between two inputs through cyclic matching, either through a cyclic loss or a proxy representation. Recent examples include the use of uniform category-level representations of objects as template 3D objects [16] or unit spheres [17]. Dense Object Nets [18] extended previous work [19] and presented a method trained with self-supervision by projecting points across multiple viewpoints of the same objects, and showing how the learned object descriptors effectively generalize over other semantically similar objects. Recently, they have been extended to learn from even less data and work with more objects through optical flow of monocular videos [20], neural radiance fields [21], and unsupervised object classification [22]. Part-based methods also often use latent representations that encode each part's appearance, shape, or pose [23, 24]. Very close to our method is [25] in that it is using cycle-consistency over transformer descriptors to produce one-shot instance segmentation.

Other methods rely on large image datasets, such as ImageNet, for pre-training of vision models as means of encoding semantic and perceptual information. Such works in point correspondence [26, 27, 28, 29] and part correspondence [23, 24] show state-of-the-art performance on several computer vision benchmarks. However, these methods often struggle with highly occluded inputs with large viewpoint variance. Caron et al. [7] show that DINO-ViT, a variant of ViT [30] trained via self-supervised knowledge distillation, can produce descriptors that contain explicit information about the underlying semantic content, with properties suitable for k-nearest neighbors (kNN) search. These properties were recently used in unison with cyclic correspondence for unparalleled co-part segmentation without fine-tuning [8]. However, the co-segmentation task does not solve our problem, since it often results in segmentation that is consistent but semantically meaningless.

The literature review leads to several conclusions: (i) for any local descriptor to be suitable for semantic correspondence, it needs to be 'aware' of the full semantic context of the object that it is part of; (ii) in matching any two objects belonging to the same semantic category, not all points will always have a true correspondence, either due to occlusion or the lack or addition of parts (such as handle on a cup, or a switch on a lamp); (iii) the currently best performing models combine semantic priors provided by pre-trained models and cyclicity.

3 Method

The one-shot semantic instance segmentation (OSIS) problem [31] is defined as finding and segmenting a previously unseen object in a novel scene, based on a single instruction example. Similarly, in this work, we aim to solve one-shot affordance part segmentation (OSAPS): given a support reference image $I_R \in \mathbb{R}^{3\times H_R\times W_R}$ and query mask region $M_Q \in \{0,1\}^{H_R\times W_R}$, the task is to find the semantically corresponding region in the target image $I_T \in \mathbb{R}^{3\times H_T\times W_T}$. The variables H_R, W_R denote the height and width of the support, while H_T, W_T denote those of the target, respectively.

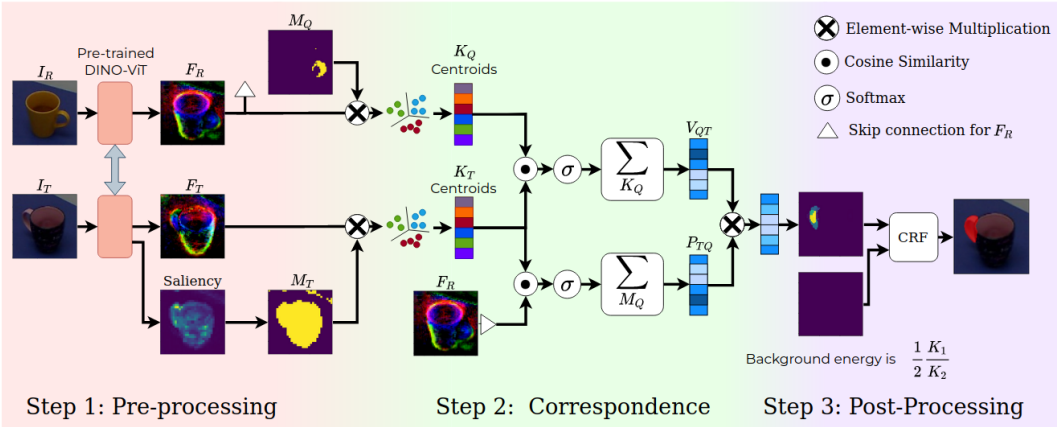

Figure 1: The overall system diagram: The support descriptors F_R belonging to the query area M_Q are clustered into K_Q query centroids. The salient target image descriptors are similarly clustered into K_T target descriptors. Cyclic correspondence is enforced by matching K_Q to K_T (forward pass), and K_T to F_T (backward pass). A score is computed for each cluster in the target. The centroid scores are mapped to the descriptors belonging to each cluster. Finally, each pixel is determined to be foreground (or not) by the CRF, which compares the scores with a baseline energy level.

3.1 Query Region Correspondence

The key takeaway from successful attempts in related works is that cyclicity plays an important role in correspondence—a matching region corresponds to the query region and vice versa (i.e., the query region corresponds to the matching region). Given the kNN-like properties of DINO-ViT, we propose to use its descriptors to find correspondences. Thus, we get descriptor images F_R and F_T corresponding to I_R and I_T, respectively. Since the descriptors retain their spatial locations, we can use the query mask M_Q and target mask M_T (e.g., one generated through saliency or user-provided) to only select the foreground pixels (Fig. 1-Step 1).

A conventional pixel correspondence method, such as [32], can then be applied to find matches between the two images. However, the method is faced with several issues due to dealing with parts instead of points: (i) occluded or missing parts/objects in the target (e.g., a cup that does not have a handle); (ii) some parts are more distinctive than others (e.g., the tip of a knife's blade versus its centre); (iii) single descriptors could be noisy and Euclidean distance between them is not sufficient for best-matching (see [18, 22]).

Here, as shown in Fig. 1-Step 2, we counteract these issues by first grouping the query descriptors into K_Q query groups and the foreground target pixels into K_T target groups. Each group is then described by its centroid descriptors F_Q and $F_{T,fg}$, respectively. Through pairwise cosine similarity, we can then find how close each of the F_Q centroids is to each $F_{T,fg}$ centroid descriptors, as $sim(F_Q, F_{T,fg}) \in \mathbb{R}^{K_Q \times K_T}$. Through the softmax operation along the target axis, we can find, for each query centroid, we determine the probability of matching each target centroid as

$$A_{QT} = \text{softmax}_{T,fg}\left(sim(F_Q, F_{T,fg})/\tau_{QT}\right), \tag{1}$$

where τ_{QT} is the softmax temperature. Similarly, we could find the backward pass, i.e., matches from the target to the support. However, in this case, we are not interested only in the matches within the query, rather in all matches anywhere in the support. With this information, we could exclude any matches that correspond to another area of the support image. Thus, our backward matching affinity is computed with the full F_R as

$$A_{TR} = \text{softmax}_{R}\left(sim(F_{T,fg}, F_R)/\tau_{TR}\right). \tag{2}$$

Summing the A_{TR} over the query pixels, gives us a probability P_{TQ} of whether each centroid corresponds to the query. Also, each centroid with $P_{TQ} > 0.5$ is likely a correspondence. Meanwhile, we also want to determine the target centroids which are the likely matches of the query centroids

in a way that allows one-to-many matching and does not only find the matches to the distinctive regions of the query. By summing over the query centroids as in Eq. (3), we get 'votes' V_{QT}. Since distinctive areas are typically smaller but produce more confident probabilities (in the affinity A_{QT}), while less distinctive ones are typically wider and have many, but lower probability, matches. This summation serves well to balance the matching scores:

$$V_{QT} = \sum_{K_Q} A_{QT} \tag{3}$$

A score for each foreground target centroid based on both passes can be obtained by $S_{T,fg} = V_{QT} \cdot P_{TQ}$. We map each descriptor to the score value that corresponds to its clustering. Since any centroid with $V_{QT} > K_Q/K_T$ is more likely than not to be a part of a match, while any centroid with $P_{TQ} > 0.5$ corresponds confidently back to the query, we determine $K_Q/2K_T$ as a score threshold for good correspondences.

Finally, as shown in Fig. 1-Step 3, to get a smoother result with the original image dimensions, the score image is passed along with the $K_Q/2K_T$ constant background energy term through a Conditional Random Field (CRF) that determines foreground and background pixels.

Design choices: Our particular choice of descriptor model is the DINO-ViT-S with patch size 8, pre-trained on ImageNet, due to its superior properties in the similar task of co-part segmentation [8]. The support and target descriptors filtering is done by applying the support query and saliency masks, respectively. Clustering is achieved via Fast K-Means [33] into an over-segmented image [1] Here, we use Fast K-Means [33] with $K_Q = K_T = 10$ [2]. We choose $\tau_{QT} = 0.2$, $\tau_{TR} = 0.02$ empirically. Having a large τ_{QT} means the forward-pass is lenient and selects many potential candidates, while the low τ_{TR} filters in only the parts of those candidates that are confidently matching to the query part rather than the rest of the reference. This setup allows the model to find one-to-many matches, unlike other correspondence works, which assume exactly one match exists [18, 20]. This comes as a trade-off of more false positives over false negatives.

3.2 Affordance Transfer

We propose one-shot affordance skill transfer by defining a stack of skills S, i.e., affordance functions $S_i(o_{aff})$ associated with manually annotated parts (and their descriptors) o_{aff} sensor inputs. Those skills would then be transferred to the corresponding parts $\hat{o}_{aff}$ of other objects by applying $S_i(\hat{o}_{aff})$. For example, a skill might be to grasp at the centroid of a graspable region, or place a ball in a containment region. In more general terms, the skill can be formulated as $S_i(\hat{o}_{aff}, o_{class}, X)$, where o_{class} is the (either categorical or latent) class of the support object and X is the robot's sensory state. The skills can be either robot-specific or robot-agnostic, since the chosen affordance representation does not make assumptions. In this work, we show this system with simple first-order affordances (grasping and containing) in simple scenes, and leave multi-object manipulations for the future.

4 Experiments

4.1 Quality of Part Transfer

Metrics: We use metrics standard in affordance learning literature [34, 35]: per-affordance class Intersection over Union (also called Jaccard index), and F_b^w-measure [36] which is a weighted version of IoU that accounts for pixel location and mask interpolation. For qualitative evaluation, we show visual comparisons.

Datasets: As our main benchmark, we use the UMD Affordance Dataset [37] due to its variety of objects, allowing us to evaluate affordance transfer in both inter- and intra-class pairs. Similarly to

[1] In theory, any similar method such as superpixelization or clustering should work as long as the final patches are dense, loosely follow the objects' geometry and produce segments smaller or equal in size to the smallest distinctive part.

[2] We found that any value larger than 3 leads to comparable results for the used datasets. However, the number of segments should be larger than the number of distinct areas of the query parts.

how works in one-shot instance transfer use modified folds of PASCAL-5^i [38] and COCO-20^i [39], we present UMD^i[3] – a one/few-shot correspondence variant of UMD, which is composed of a single instance of each object in the dataset with both RGB image and affordance ground truth annotation. The original UMD annotations are kept to highlight the difficulties of semantic transfer, as no two human annotations are the same. The classes include common objects such as bowls, ladles, scissors, and knives, with manually labelled annotations of grasp, scoop, wrap-grasp, support, contain, cut, and pound affordances.

Experimental Setup: For each object in UMD^i, we attempt to transfer its ground truth part masks to each other object of the same class (*intra-class*) and to each other object of classes that possess the same affordance (*inter-class*) from the dataset. The quantitative metrics are averaged over each affordance type. The qualitative comparison is shown as well.

Baselines: As an upper baseline, we include fully supervised model reported scores [34, 35] on the UMD dataset (standard split for intra- and novel split for inter-class results). These are not directly comparable with our method, since they could learn the way affordances are labelled within the dataset through supervision and are evaluated on the test subset of UMD, instead of UMD^i. As baselines, we have included two variants of the current SotA in one-shot instance transfer–BAM [40]. The unsupervised co-segmentation method [8], which inspired our work, was included by taking all parts that have at least 50% overlap with the ground truth parts of the support image and using their correspondences as the 'estimated parts' in the target image. It is given the support and target images, and searches for between 2 and 10 parts with K-Means elbow point as in [8].

Finally, to showcase the difficulty of semantic labelling we also included a 'human level': a person is shown two objects from the same class, they are given the task of one-shot transferring labels by observing the ground truth annotation of one image and annotating the second. This is done once for each UMD^i object as the target, after which the metric score is scaled by the number of objects belonging to that class before computing the per-affordance means.

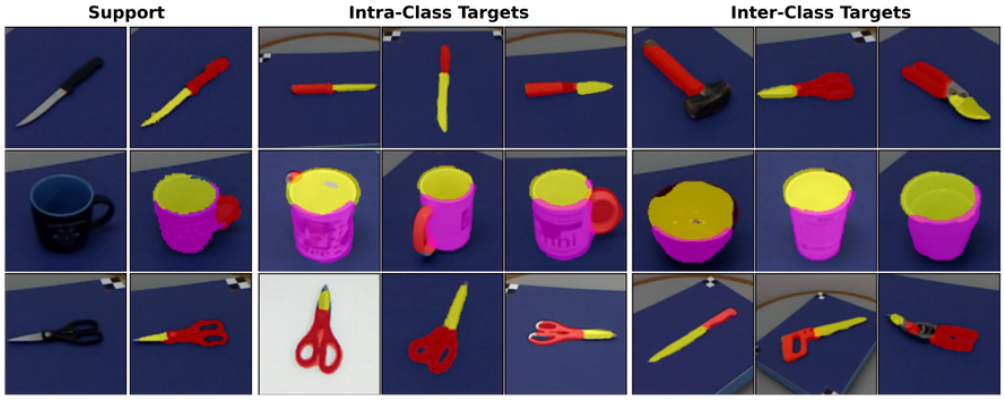

Figure 2: The annotated affordance regions from each support image are transferred with AffCorrs to intra- and inter-class targets on the same row. Colors represent grasp, cut, contain, wrap-grasp.

Results: A visualization of parts transferred with AffCorrs in Figure 2 shows that the masks are relatively robust to viewpoint variance (see mugs), and missing correspondences (see knife-to-hammer, and mug-to-cup), while being surprisingly capable of transferring affordance across dissimilar regions. In Figure 3 we show a comparison with the unsupervised baselines. Both the qualitative and quantitative comparisons (in Tables 1 and 2) affirm that AffCorrs performs better on UMD^i[4]. The BAM baseline under-performs, likely due to being tailored for whole object instance transfer rather than parts. Meanwhile, the co-part segmentation, which uses the same saliency masks and backbone as AffCorrs, appears to often ignore the foreground when it doesn't deem it common enough across the objects, and produce parts that don't align with what we would consider semantically significant.

[3]The dataset and code will be released upon acceptance.

[4]Further results can be found in the supplementary material.

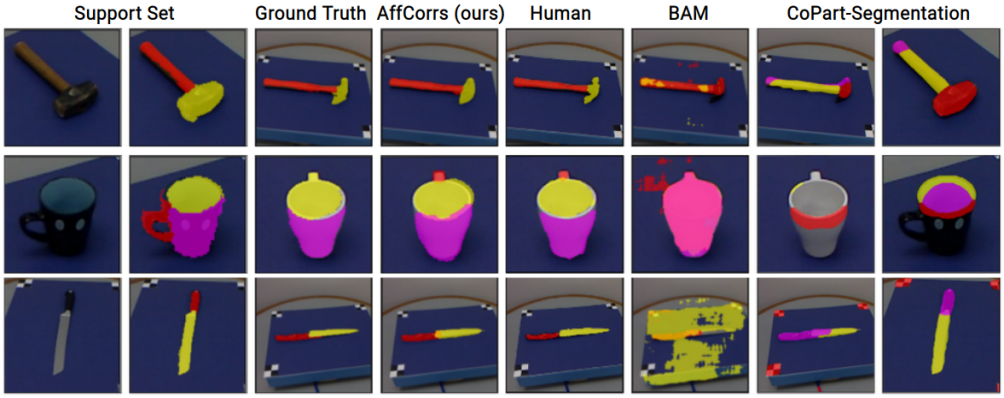

Figure 3: Visual comparison of the segmentation produced by various methods. Colours represent grasp, cut, contain, wrap-grasp. Note that Co-part segmentation has a separate colouring.

	Grasp		Cut		Scoop		Contain		Wrap-grasp		Pound		Support	
	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w
Supervised														
ResNet [34]	**0.71**	-	**0.79**	-	**0.86**	-	**0.86**	-	**0.84**	-	**0.72**	-	**0.55**	-
ADNet [35]	-	**0.73**	-	0.72	-	**0.80**	-	**0.85**	-	0.81	-	**0.87**	-	0.76
AffNet [3]	-	**0.73**	-	**0.81**	-	0.76	-	0.83	-	**0.82**	-	0.79	-	**0.84**
Unsupervised / One-Shot Transfer														
BAM-ResNet [40]	0.26	0.26	0.28	0.23	0.52	0.57	0.57	0.60	0.42	0.45	0.45	0.50	0.43	0.60
BAM-VGG [40]	0.15	0.17	0.17	0.13	0.43	0.45	0.56	0.59	0.41	0.45	0.39	0.44	0.27	0.41
DINO-ViT [8]	0.45	0.51	0.57	0.64	0.61	0.64	0.42	0.48	0.53	0.62	0.66	0.76	0.66	0.75
AffCorrs (ours)	**0.55**	**0.65**	**0.72**	**0.81**	**0.73**	**0.81**	**0.82**	**0.87**	**0.83**	**0.89**	**0.78**	**0.87**	**0.82**	**0.87**
Human level	0.59	0.79	0.64	0.82	0.66	0.83	0.72	0.79	0.73	0.74	0.74	0.74	0.74	0.75

Table 1: Comparison of per-affordance metrics on intra-class pairs.

	Grasp		Cut		Scoop		Contain		Wrap-grasp		Pound		Support	
	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w	IoU	F_β^w
Supervised														
ResNet [34]	**0.33**	-	**0.51**	-	**0.69**	-	**0.52**	-	**0.85**	-	**0.09**	-	**0.51**	-
Unsupervised / One-Shot Transfer														
BAM-ResNet [40]	0.22	0.25	0.22	0.25	0.20	0.21	0.51	0.54	0.17	0.18	0.15	0.16	0.12	0.13
BAM-VGG [40]	0.13	0.15	0.13	0.14	0.17	0.18	0.50	0.52	0.16	0.18	0.13	0.15	0.05	0.05
DINO-ViT [8]	**0.39**	**0.45**	0.50	**0.57**	0.58	0.60	0.30	0.34	0.56	0.64	0.66	**0.75**	0.68	0.76
AffCorrs (ours)	**0.39**	0.41	**0.51**	0.50	**0.62**	**0.65**	**0.71**	**0.75**	**0.83**	**0.87**	**0.72**	0.73	**0.82**	**0.79**

Table 2: Comparison of per-affordance metrics on inter-class pairs.

4.2 Affordance Transfer in the Real World

Experimental Setup: To showcase the application of AffCorrs to affordance transfer and evaluate the current limitations in realistic scenes, we present the following evaluation setup: we use a Franka Emika Panda robot with an arm-mounted RGB-D camera. As a query, we use a single image of a screwdriver toy with the annotated grasping area, and a mug with the annotated containment. The query is one-shot transferred to the robot's unseen environment. Finally, the robot attempts to use the affordance skill.

We test the method in two settings (See Figure 4) – single object and multiple objects. Single object scenes contain one object that belongs to the same class as the query (intra-class), or an object from a different class (inter-class) that has an affordance equivalent to the queried one (e.g., graspable part). In the scenes containing multiple objects, there are several objects that are 'true' correspondences,

along with distractor objects. In both cases, the background was varied with different surfaces (table, carton, textile, cloths). The objects used are relatively spaced out, with occasional occlusion. Ten examples of either setting were used for the following robot experiments.

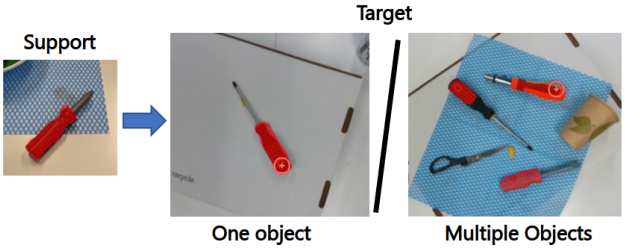

Figure 4: Examples of robot test scenes - either with one object or multiple. The red area denotes correspondences, the circle - the next object selected for manipulation based on the highest point.

Two simple first-order affordance skills are shown: grasping and containment. More complex multi-object interactions are left for future work. The grasping skill is defined as picking at the 3D centroid of a part, with grasp orientation along the largest PCA axis in XY-space (i.e., we assume a top grasp). The containment skill is defined as opening the robot gripper above the XY centroid of the affordance region. Each affordance is attempted ten times for each setting. An attempt is deemed successful if the robot successfully uses the affordance of all objects that correspond to the query, e.g., grasps all objects with tool handles *and* doesn't attempt to grasp a non-corresponding object.

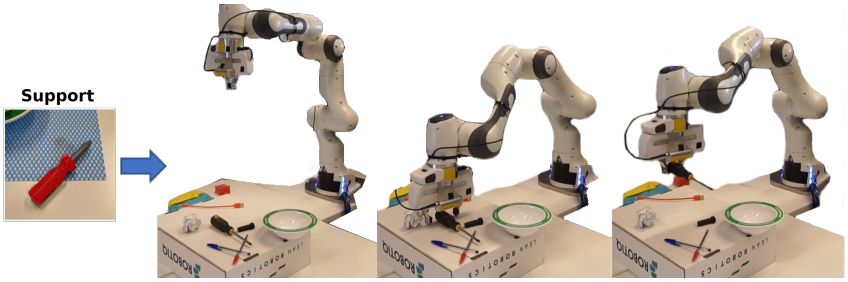

Figure 5: Example of the robot finding and using the handle-grasp affordance.

Results: An example grasp in multi-object scene is shown in Figure 5. In the one-object setting, we observed a 100% success rate for both grasping and containment. In the multi-object setting, the grasping success rate dropped to 70% and the containment – to 80%. As noted by previous OSIS approaches [31, 25], AffCorrs too shows that cluttered scenes are more challenging, which underscores the importance of extending the model in order to process such scenes robustly.

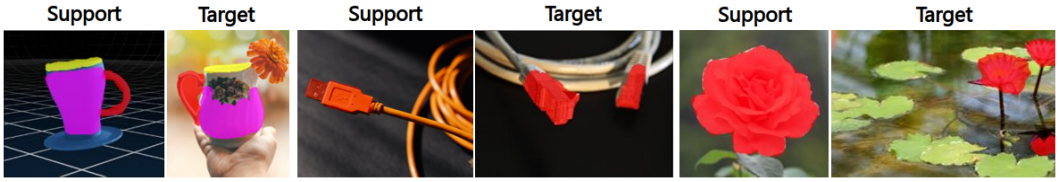

Figure 6: Several examples of one-shot transfer, from left to right: Transfer (i) from a mug simulated in Isaac Sim to a real jug; (ii) from a USB cable to an Ethernet cable; (iii) from a rose to water loti. Note that the reflection of the flower was marked as corresponding as well.

4.3 Transfer across Less Common Objects

Finally, we briefly motivate the practical usefulness of the one-shot property of AffCorrs over supervised methods: AffCorrs can produce good semantic part correspondences across less common objects, allowing it to work with very specific affordances that likely are not represented in any big

dataset. Moreover, transferring a skill from simulation to reality could be made easy by using such an affordance representation; see examples in Figure 6.

5 Limitations

We highlight several important limitations of the current method (see Figure 7): AffCorrs searches for correspondences with probability-based cycle-consistency, which means that it can find correspondences between vastly different objects provided that they have similar descriptors. This is in the method's favour when the support image is suitable - e.g., when transferring from a screwdriver to a hammer's handle, but not as much when the support image is different, such as a bowl to a hammer. We have seen that the descriptor similarity alone may not be enough to always prevent incorrect matches from appearing, which limits the current model to simpler scenes without severe clutter. This is also limiting the method's ability to perform well in clutter. Meanwhile, the alternative of having more conservative matching would limit the transfer-ability across inter-class pairs.

Another issue stems from the descriptors themselves – the DINO-ViT descriptors confuse between an object's texture (e.g., a print of a dog on a mug) and an actual object (e.g., a real dog). This would mean that two same-class objects with and without textures sometimes result in different parts. The transformer's positional encoding makes the model biased toward picking a correspondence that is similarly located within the image instead of the actual correspondence. This limitation could potentially be addressed by performing flipping and cropping augmentations before the clustering steps. In terms of affordances, this work shows some simple affordance interactions, however the method could, in theory, be applied to multi-object interactions. The limitations of that direction are yet to be assessed.

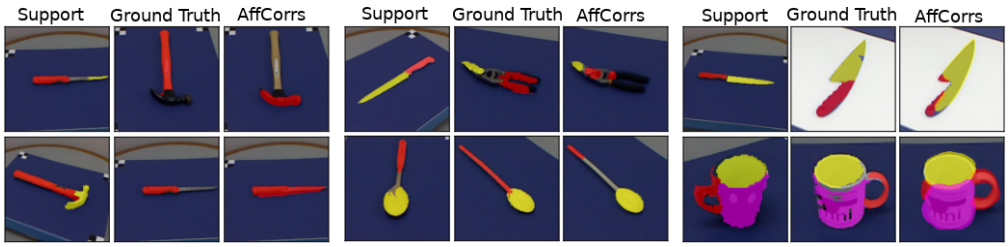

Figure 7: Six examples of model failures: some failure can likely be explained by the colour and shape of the parts shown on the top row of examples, e.g., a black knife handle mismatched to a black hammer. AffCorrs also fails due to the inherited positional encoding, observed in the bottom row examples.

6 Conclusion and Future Work

In this work, we showed how pre-trained DINO-ViT's descriptors can be adapted for part-based transfer of affordances. We have shown that AffCorrs is better suited for this task than the current best one-shot instance segmentation baseline and has an impressive cross-object transfer of part segmentation. However, it's limitations in working with clutter need to be addressed before any real-world application. Potential candidates to solve this problem include using an object detection model (such as DetCo [41]), more sophisticated matching, or some form of unsupervised latent classification. Solving this issue would make the method suitable for more complex dataset benchmarks, such as the IIT's affordance dataset [10]. While this work is the first tackling the one-shot affordance transfer problem, it opens the door for many future directions such as learning to both discover and use affordances from one-shot observations by looking at how people interact with objects; using affordance regions for transfer of more complex multi-object interactions; assisting teleoperation using affordance-guided object manipulation rather than hand-to-robot movement transfer.

References

[1] J. J. Gibson. The theory of affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Erlbaum, 1977.

[2] J. Borja-Diaz, O. Mees, G. Kalweit, L. Hermann, J. Boedecker, and W. Burgard. Affordance Learning from Play for Sample-Efficient Policy Learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[3] T.-T. Do, A. Nguyen, and I. Reid. AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] A. Nguyen, D. Kanoulas, D. Caldwell, and N. Tsagarakis. Detecting Object Affordances with Convolutional Neural Networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016. doi:10.1109/IROS.2016.7759429.

[5] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3): 42–62, 2022. doi:10.1109/MSP.2021.3134634.

[6] J. M. Wong. Towards Lifelong Self-Supervision: A Deep Learning Direction for Robotics. *CoRR*, abs/1611.00201, 2016.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[8] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep ViT Features as Dense Visual Descriptors. *arXiv preprint arXiv:2112.05814*, 2021.

[9] T. Lüddecke and F. Wörgötter. Learning to Segment Affordances. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 769–776, 2017. doi:10.1109/ICCVW.2017.96.

[10] 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, 2017. doi:10.1109/IROS.2017.8206484.

[11] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis. Translating Videos to Commands for Robotic Manipulation with Deep Recurrent Neural Networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3782–3788, 2018. doi:10.1109/ICRA.2018.8460857.

[12] T. Mar, V. Tikhanoff, G. Metta, and L. Natale. Self-supervised learning of grasp dependent tool affordances on the iCub Humanoid robot. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3200–3206, 2015.

[13] T. Mar, V. Tikhanoff, G. Metta, and L. Natale. Self-supervised learning of tool affordances from 3D tool representation through parallel SOM mapping. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 894–901, 2017. doi:10.1109/ICRA.2017.7989110.

[14] P. Abelha, F. Guerin, and M. Schoeler. A model-based approach to finding substitute tools in 3D vision data. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2471–2478, 2016. doi:10.1109/ICRA.2016.7487400.

[15] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao. One-Shot Affordance Detection. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[16] B. Wen, W. Lian, K. Bekris, and S. Schaal. CaTGrasp: Learning Category-Level Task-Relevant Grasping in Clutter from Simulation. *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[17] A.-C. Cheng, X. Li, M. Sun, M.-H. Yang, and S. Liu. Learning 3D Dense Correspondence via Canonical Point Autoencoder. In *Advances in Neural Information Processing Systems*, 2021.

[18] P. Florence, L. Manuelli, and R. Tedrake. Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation. *Proceedings of the 1st Conference on Robot Learning (CoRL)*, 2018.

[19] T. Schmidt, R. A. Newcombe, and D. Fox. Self-Supervised Visual Descriptor Learning for Dense Correspondence. *IEEE Robotics and Automation Letters*, 2:420–427, 2017.

[20] U. Deekshith., N. Gajjar., M. Schwarz., and S. Behnke. Visual Descriptor Learning from Monocular Video. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 444–451. INSTICC, SciTePress, 2020. ISBN 978-989-758-402-2. doi:10.5220/0008989304440451.

[21] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola. NeRF-Supervision: Learning dense object descriptors from neural radiance fields. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[22] D. Hadjivelichkov and D. Kanoulas. Fully Self-Supervised Class Awareness in Dense Object Descriptors. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning (CoRL)*, volume 164 of *Proceedings of Machine Learning Research*, pages 1522–1531. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/hadjivelichkov22a.html.

[23] W.-C. Hung, V. Jampani, S. Liu, P. Molchanov, M.-H. Yang, and J. Kautz. SCOPS: Self-Supervised Co-Part Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Q. Gao, B. Wang, L. Liu, and B. Chen. Unsupervised Co-part Segmentation through Assembly. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 3576–3586. PMLR, 2021.

[25] G. Zhang, G. Kang, Y. Yang, and Y. Wei. Few-Shot Segmentation via Cycle-Consistent Transformer. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 21984–21996, 2021.

[26] U. Gaur and B. S. Manjunath. Weakly Supervised Manifold Learning for Dense Semantic Object Correspondence. In *International Conference on Computer Vision (ICCV)*, pages 1744–1752, 2017. doi:10.1109/ICCV.2017.192.

[27] J. Lee, D. Kim, W. Lee, J. Ponce, and B. Ham. Learning Semantic Correspondence Exploiting an Object-level Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[28] J. Lee, D. Kim, J. Ponce, and B. Ham. SFNet: Learning Object-aware Semantic Flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] S. Choudhury, I. Laina, C. Rupprecht, and A. Vedaldi. Unsupervised Part Discovery from Contrastive Reconstruction. In *Advances in Neural Information Processing Systems*, 2021.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.

[31] C. Michaelis, I. Ustyuzhaninov, M. Bethge, and A. S. Ecker. One-Shot Instance Segmentation. *CoRR*, abs/1811.11507, 2018. doi:10.48550/ARXIV.1811.11507.

[32] T. Xiao, S. Liu, S. De Mello, Z. Yu, J. Kautz, and M.-H. Yang. Learning Contrastive Representation for Semantic Correspondence. *International Journal for Computer Vision (IJCV*, 130(5):1293–1309, 2022. ISSN 0920-5691. doi:10.1007/s11263-022-01602-y.

[33] G. Hamerly and J. Drake. *Accelerating Lloyd's Algorithm for K-Means Clustering*, pages 41–78. 2015. doi:10.1007/978-3-319-09259-1_2.

[34] J. Sawatzky, A. Srikantha, and J. Gall. Weakly Supervised Affordance Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] K. Chaudhary, K. Okada, M. Inaba, and X. Chen. Predicting Part Affordances of Objects Using Two-Stream Fully Convolutional Network with Multimodal Inputs. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3096–3101, 2018. doi:10.1109/IROS.2018.8593617.

[36] R. Margolin, L. Zelnik-Manor, and A. Tal. How to Evaluate Foreground Maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014. doi:10.1109/CVPR.2014.39.

[37] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance Detection of Tool Parts from Geometric Features. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[38] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots. One-Shot Learning for Semantic Segmentation. In *British Machine Vision Conference (BMVC)*. BMVA Press, 2017.

[39] K. Nguyen and S. Todorovic. Feature Weighting and Boosting for Few-Shot Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 622–631. IEEE, 2019. doi:10.1109/ICCV.2019.00071.

[40] C. Lang, G. Cheng, B. Tu, and J. Han. Learning What Not to Segment: A New Perspective on Few-Shot Segmentation. 2022. doi:10.48550/arXiv.2203.07615.

[41] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo. DetCo: Unsupervised Contrastive Learning for Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8372–8381. IEEE, 2021. doi:10.1109/ICCV48922.2021.00828.