

One-Shot Transfer of Affordance Regions? AffCorrs! Supplementary Meterials

Denis Hadjivelichkov *
 Centre for Artificial Intelligence
 University College London

Sicelukwanda Zwane
 Centre for Artificial Intelligence
 University College London

Marc Deisenroth
 Centre for Artificial Intelligence
 University College London

Lourdes Agapito
 Department of Computer Science
 University College London

Dimitrios Kanoulas
 Department of Computer Science
 University College London

A Appendix: AffCorrs Variants Ablation

In Table 1 we compare the performance of the proposed model when the cyclicity is broken, i.e. only one of the correspondence directions is kept active. The variants using only either P_{TQ} or V_{QT} in the calculation of the scores (S_{T,f_g} in method), while the other is set to 1. The threshold used for the CRF background energy is not calculated, but instead chosen as the best performing threshold from a parameter sweep. The rest of the model is kept the same. The performance metrics are calculated on the intra-class UMDⁱ task. We observe that indeed both branches alone perform worse than when together, but also that they are competitive with the best performing unsupervised baseline.

	Grasp		Cut		Scoop		Contain		Wrap-grasp		Pound		Support	
	IoU	F_β^w												
DINO-ViT	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 Variants														
P_{TQ} and V_{QT}	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
P_{QT} only	0.45	0.57	0.53	0.67	0.61	0.71	0.68	0.78	0.70	0.84	0.66	0.78	0.68	0.77
V_{TQ} only	0.45	0.44	0.62	0.62	0.65	0.64	0.61	0.61	0.59	0.59	0.73	0.74	0.73	0.73

Table 1: Comparison of different AffCorrs variants

B Affordance Transfer Comparison

	Grasp		Contain	
	Single Object	Multiple Objects	Single Object	Multiple Objects
AffCorrs	100%	70%	100%	80%
BAM ResNet	20%	0%	20%	0%
BAM VGG	20%	0%	30%	0%
DINO-ViT	80%	40%	20%	0%

Table 2: Comparison of the grasping success rates

*Corresponding authors are {dennis.hadjivelichkov, dimitrios.kanoulas}@ucl.ac.uk

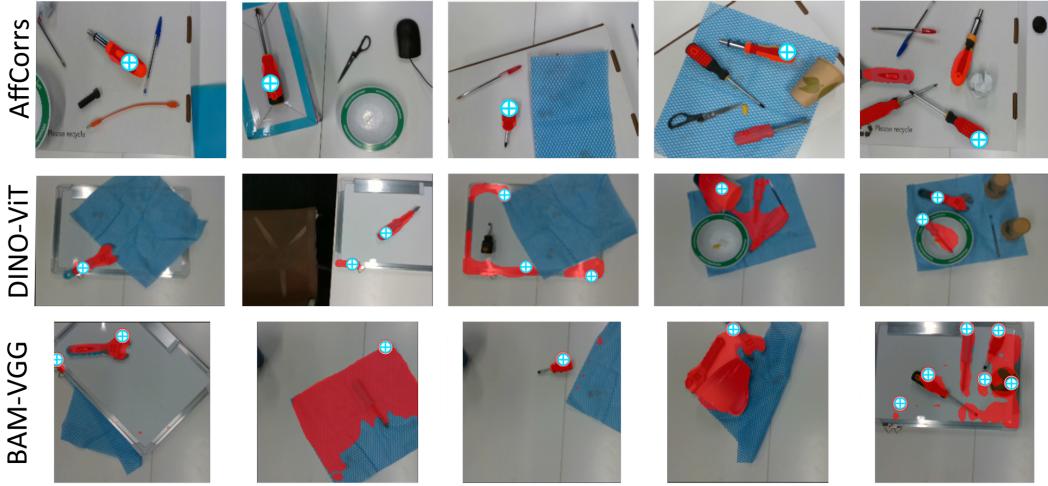


Figure 1: Comparison between the grasping points generated by the three methods, shown as blue crosshairs with corresponding areas highlighted in red. Note that the AffCorrs row only shows grasping points for the next grasp to be executed. Some baseline grasps were successful despite the correspondence being wrong (e.g., DINO-ViT, first column)

The baselines are used to compare the affordance transfer success rates - 10 trials are done in single- and multiple- object settings, repeated for each affordance. The results are shown in Table 2, with grasping point examples shown in Figure 1.

With both skills, the BAM baseline fails to produce good part correspondences, and often estimates the full image as a correspondence. The DINO-ViT Co-part segmentation baseline estimates the common parts between the support and the target, decides which parts are part of the support (estimating the support by selecting the parts that have big overlap with the support mask, and aggregating them together), and finally selects the parts that correspond to them in the target. In the single object grasping setting, while the selected areas are often observed to be ‘wrong’, they are good enough to grasp the object with the same skill. In the multiple object case, the co-part segmentation estimate often (i) does not separate the support into ‘correct’ parts and (ii) confuses distractor objects with the query. When dealing with mugs, we observed a significant drop in performance even in the single object, likely explained by the significantly different top-down viewpoint.

C Objects used in Robot Experiments

The objects shown in Figure 2 were used for the robot experiments, both for single- and multi-object settings. Distractor objects are not shown.

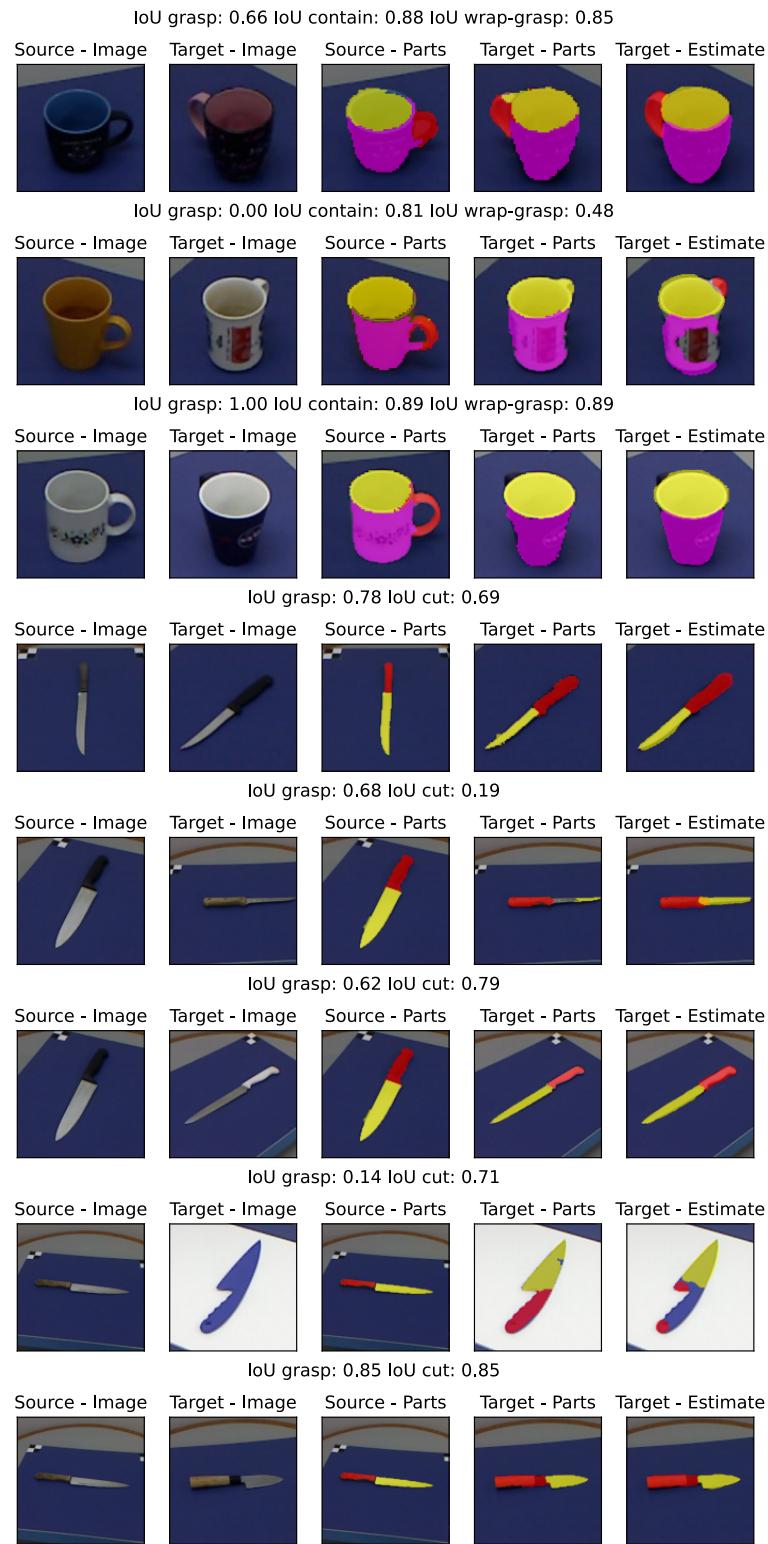
D Intra-Class Qualitative Results

In the following figures, the support set consists of the source - image and its annotated source - parts. The target image and its ground truth parts are shown. Finally, the AffCorrs output is denoted as “Target - Estimate”. Each example row also shows the corresponding affordance IoU scores above it. These supplementary results show both the strengths and the weaknesses of the model in dealing with different shapes and different textures.



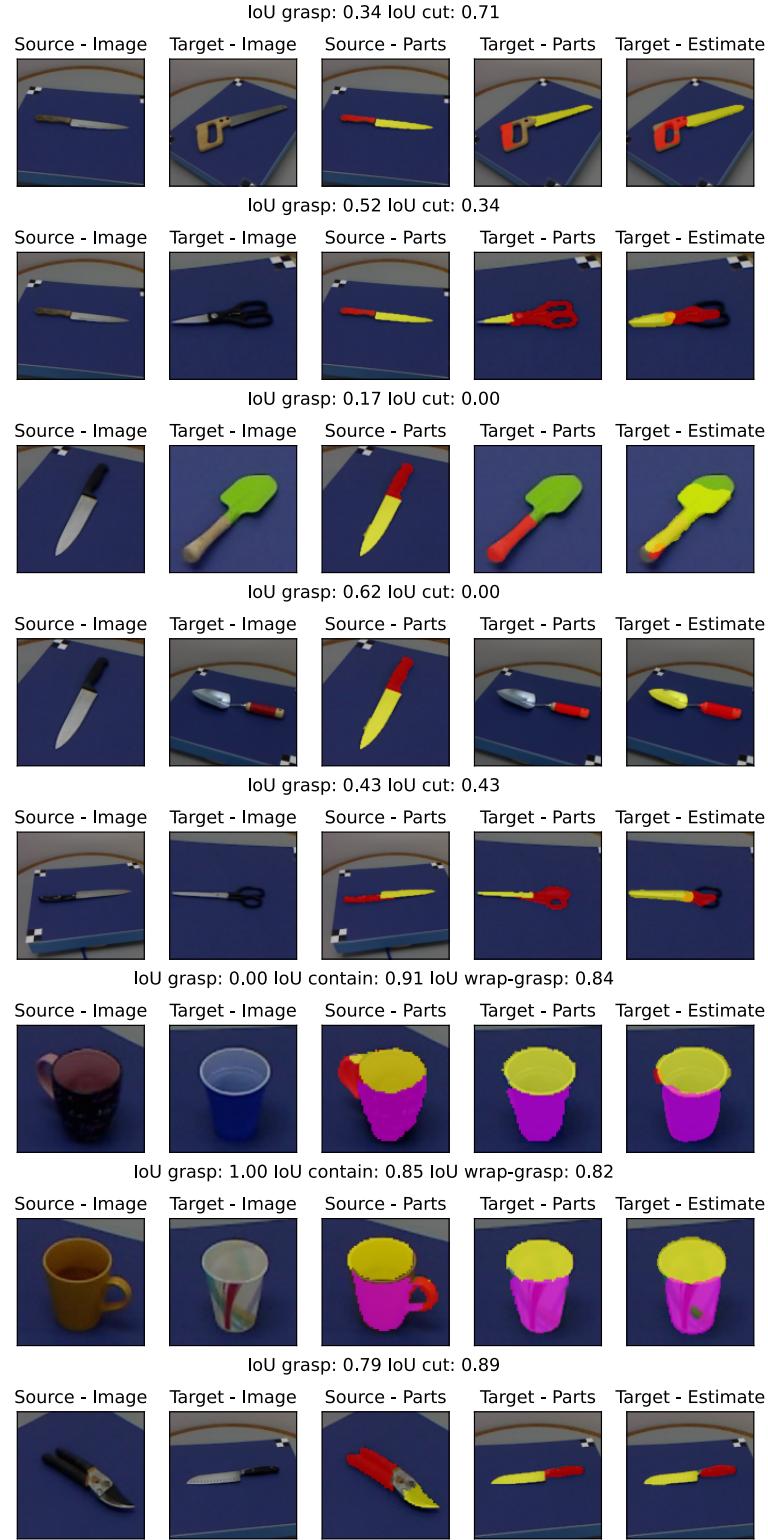
Figure 2: Objects used in robot experiments: top row, containment affordance; bottom row, tool grasp affordance.

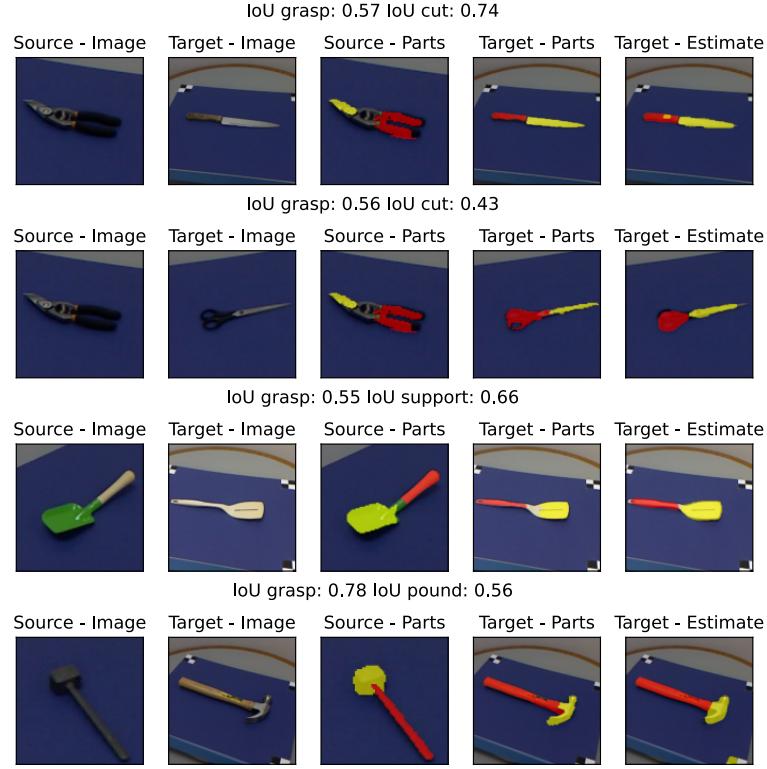




E Inter-Class Qualitative Results

The inter-class figures follow the same pattern as intra-class: Support image, Target Image, Support query, Target Ground Truth, and finally - AffCorrs output.





F Flow-Based Methods



Figure 3: Flow-based segment transfer on UMD dataset. WarpC-SemanticGLUNet shows the support mask transferred onto the respective target, while the Warped Support shows the warping applied onto the support image.

We use the recent WarpC-Semantic GLUNet [1] based from the UCN family of flow-based methods. We show that the flow-based dense correspondence method seems to do well when dealing with very similar objects, but they do not perform that well when the objects are oriented or look differently despite belonging to the same class (Figure 3). Note that knives and trowels are not categories present in the model’s dataset (Spair-71k), but humans that carry similar objects are.

References

- [1] P. Truong, M. Danelljan, F. Yu, and L. V. Gool. Warp Consistency for Unsupervised Learning of Dense Correspondences. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021. URL <https://arxiv.org/abs/2104.03308>.