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 ($S_{T,fg}$ 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 UMD 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.

<table>
<thead>
<tr>
<th></th>
<th>Grasp</th>
<th>Cut</th>
<th>Scoop</th>
<th>Contain</th>
<th>Wrap-grasp</th>
<th>Pound</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IoU</td>
<td>$F_g^w$</td>
<td>IoU</td>
<td>$F_g^w$</td>
<td>IoU</td>
<td>$F_g^w$</td>
<td>IoU</td>
</tr>
<tr>
<td>DINO-ViT</td>
<td>0.45</td>
<td>0.51</td>
<td>0.57</td>
<td>0.64</td>
<td>0.61</td>
<td>0.64</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>AffCorrs Variants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{TQ}$ and $V_{QT}$</td>
<td>0.55</td>
<td>0.65</td>
<td>0.72</td>
<td>0.81</td>
<td>0.73</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>$P_{TQ}$ only</td>
<td>0.45</td>
<td>0.57</td>
<td>0.53</td>
<td>0.67</td>
<td>0.61</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>$V_{QT}$ only</td>
<td>0.45</td>
<td>0.44</td>
<td>0.62</td>
<td>0.62</td>
<td>0.65</td>
<td>0.64</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 1: Comparison of different AffCorrs variants

B Affordance Transfer Comparison

<table>
<thead>
<tr>
<th></th>
<th>Grasp</th>
<th>Contain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single Object</td>
<td>Multiple Objects</td>
</tr>
<tr>
<td>AffCorrs</td>
<td>100%</td>
<td>70%</td>
</tr>
<tr>
<td>BAM ResNet</td>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td>BAM VGG</td>
<td>20%</td>
<td>0%</td>
</tr>
<tr>
<td>DINO-ViT</td>
<td>80%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the grasping success rates

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

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.
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.

- IoU grasp: 0.34 IoU cut: 0.71
  ![Image](image1)

- IoU grasp: 0.52 IoU cut: 0.34
  ![Image](image2)

- IoU grasp: 0.17 IoU cut: 0.00
  ![Image](image3)

- IoU grasp: 0.62 IoU cut: 0.00
  ![Image](image4)

- IoU grasp: 0.43 IoU cut: 0.43
  ![Image](image5)

- IoU grasp: 0.00 IoU contain: 0.91 IoU wrap-grasp: 0.84
  ![Image](image6)

- IoU grasp: 1.00 IoU contain: 0.85 IoU wrap-grasp: 0.82
  ![Image](image7)

- IoU grasp: 0.79 IoU cut: 0.89
  ![Image](image8)
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.

F Flow-Based Methods
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