One-Shot Transfer of Affordance Regions? AffCorrs!

Denis Hadjivelichkov, Sicelukwanda Zwane, Marc P. Deisenroth, Lourdes Agapito, Dimitrios Kanoulas

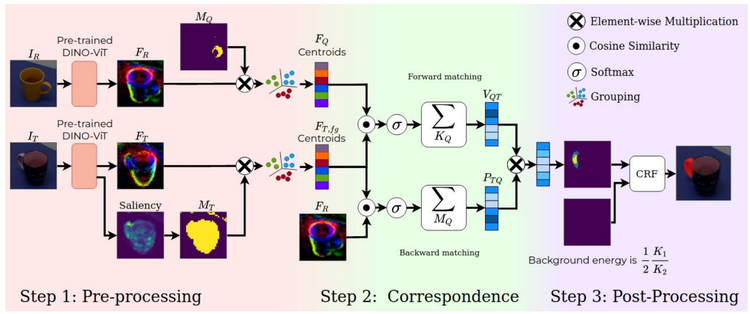
UCL

Problem

- How can robots use novel objects?
- Can we use known affordance masks to find corresponding affordance regions in unseen scenes?

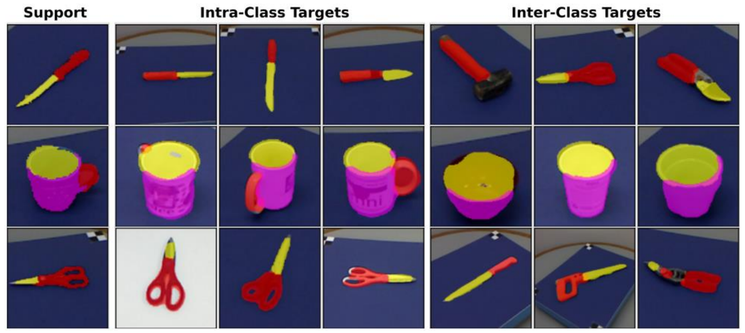
Given a **single** reference image and its annotation, we want to find its region correspondence in the target image.

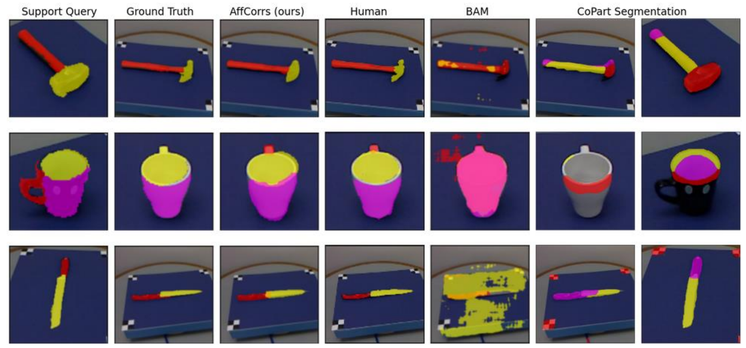
Method

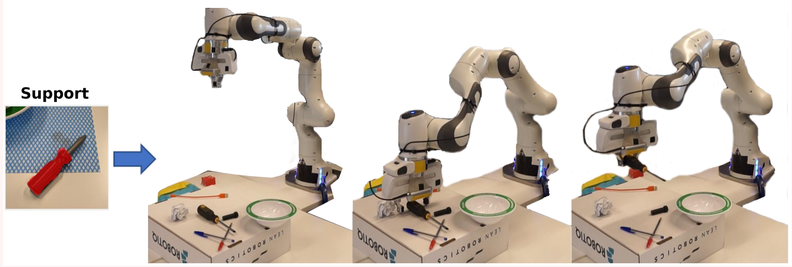


1. We utilize pre-trained DINO-ViT model to find descriptors retaining semantic context.
2. The descriptors of the query and target images are grouped into clusters.
3. Each group in either image 'votes' for the groups in the other image, ensuring cyclic bi-directional matching.
4. Finally, descriptor groups in the target image are scored, and a smooth mask is produced via CRF.

Results







	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

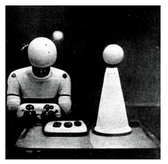
Table 1: Comparison of per-affordance metrics on intra-class pairs.

	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**

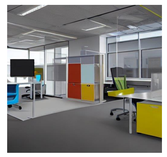
Table 2: Comparison of per-affordance metrics on inter-class pairs.

What's Next?

One-Shot Imitation



Assisted Teleoperation



Scene Understanding

Depth Perception

Limitations

AffCorrs struggles to find good matches in severe clutter. Moreover, the model confuses between an object's texture and an actual object. The transformer's positional encoding biases the model toward matching similar locations instead of the actual correspondence.

Corresponding author:
dennis.hadjivelichkov@ucl.ac.uk