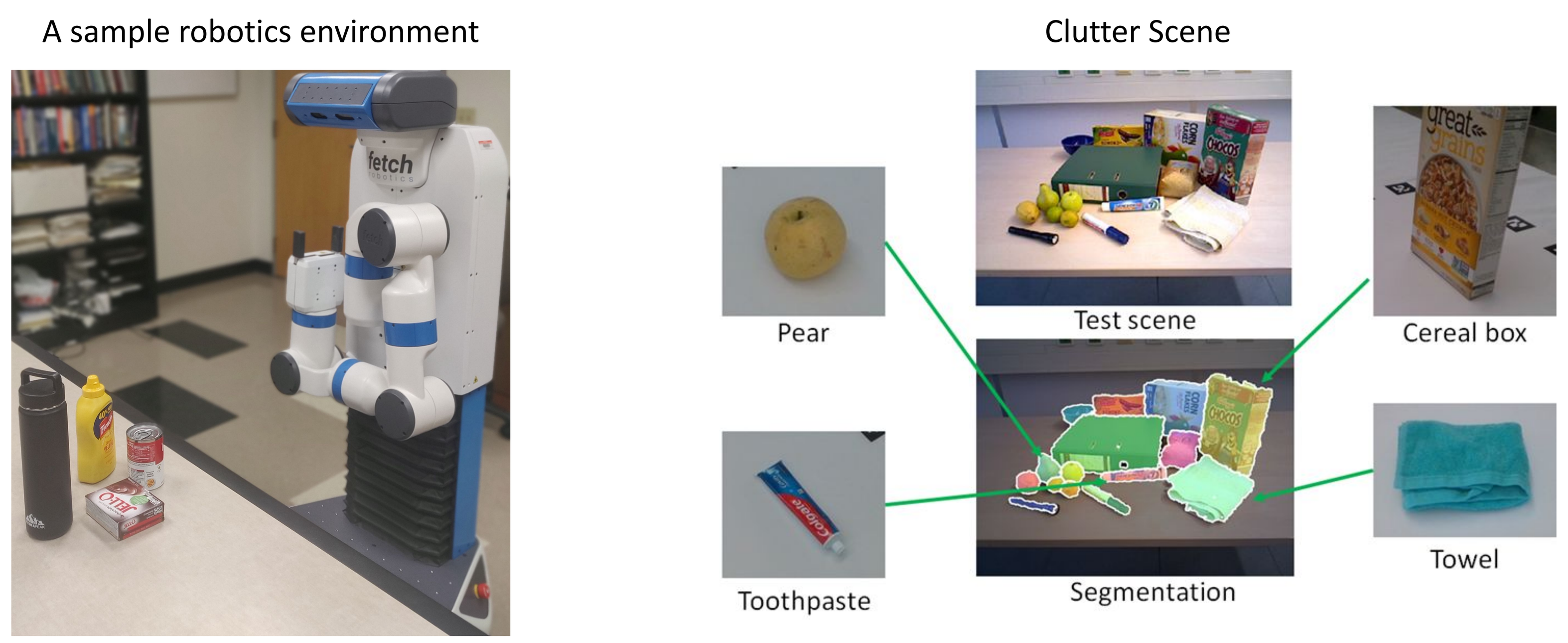
FewSOL: A Dataset for Few-Shot Object Learning in Robotic Environments

Jishnu Jaykumar P [1] Yu-Wei Chao [2] Yu Xiang [1]

[1]The University of Texas at Dallas [2]NVIDIA

Motivation



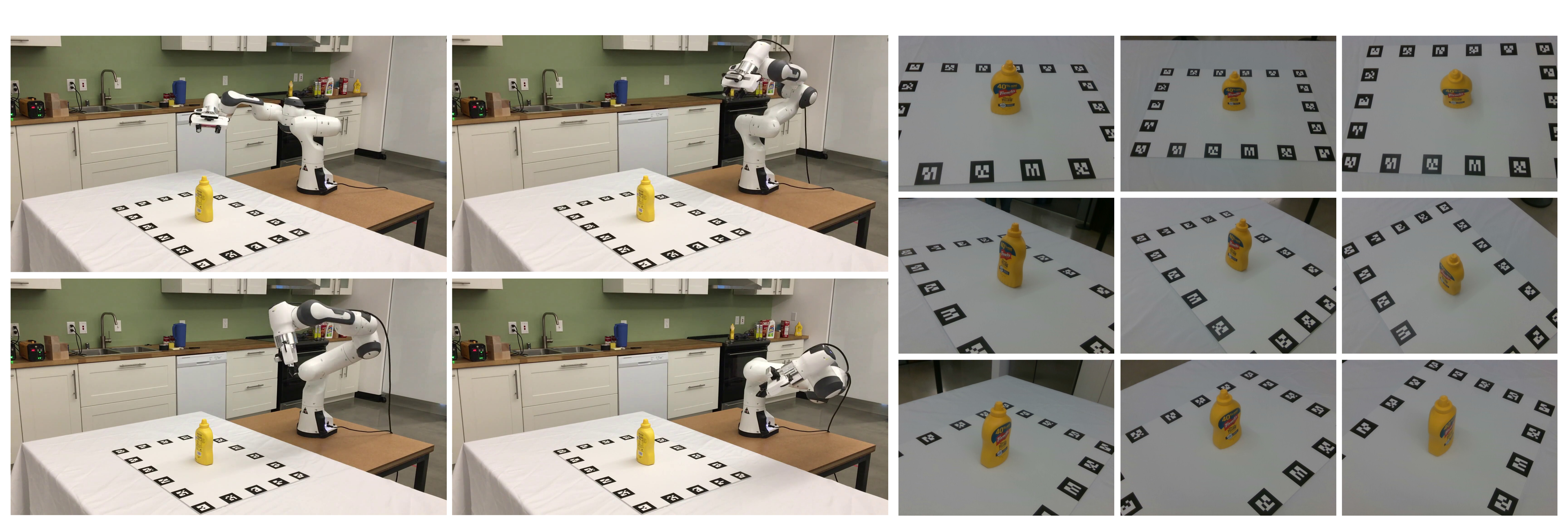
Goal: A robot should identity various (daily) objects in clutter scenes
Our approach: Object Classification using Few-Shot Learning

Related versus Our Dataset

Dataset	Class type	#classes	Image Type	#images_per_class	Annotations
Omniglot	Characters	1623	RGB	20	class label
mini-ImageNet	WordNet synsets	100	RGB	600	class label
ILSVRC-2012	WordNet synsets	1000	RGB	≈8,004	class label
Aircraft	Aircraft	100	RGB	100	class label
CUB-200-2011	Birds	200	RGB	≈59	class label
Describable Texture	Textures	47	RGB	120	class label
Quick Draw	Drawings	345	RGB	≈146,164	class label
Fungi	Fungal species	1394	RGB	≈65	class label
VGG Flower	Flowers	102	RGB	≈81	class label
Traffic Signs	Traffic signs	43	RGB	≈912	class label
MSCOCO	Internet objects	80	RGB	≈10,751	class label, segmentation
Ours (real + synthetic)	**Daily objects**	**198 + 125**	**RGB-D**	**≈27 + 10,234**	**class label, segmentation object pose and attribute**

Comparison of our dataset with other datasets for few-shot learning in the literature. Our dataset contains daily objects in robot manipulation settings with both real and synthetic images and additional annotations other than object class label.

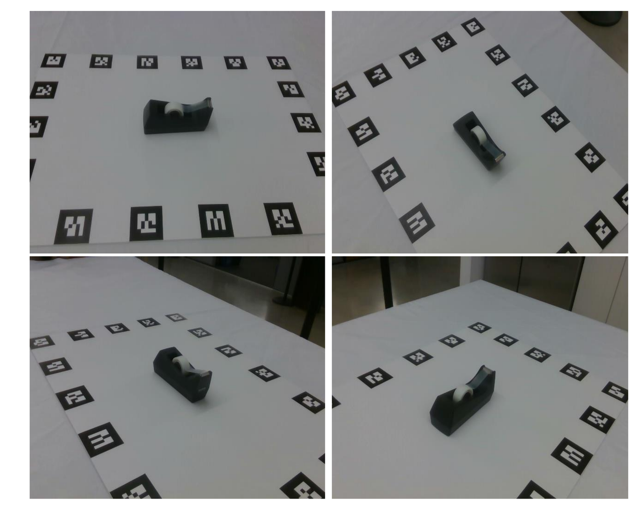
Dataset Construction: 1. Data Capture in the Real World



(a) (b)

(a) Our data capture system with a Franka Emika Panda arm. (b) 9 images of a mustard bottle from different views captured in our dataset.

Dataset Construction: 2. Data Annotation



1. What is the name of the object in these images?
tape and tape holder, tape, support and adhesive tape

2. What is the category of the object in these images?
stationary, stationery / adhesive tape

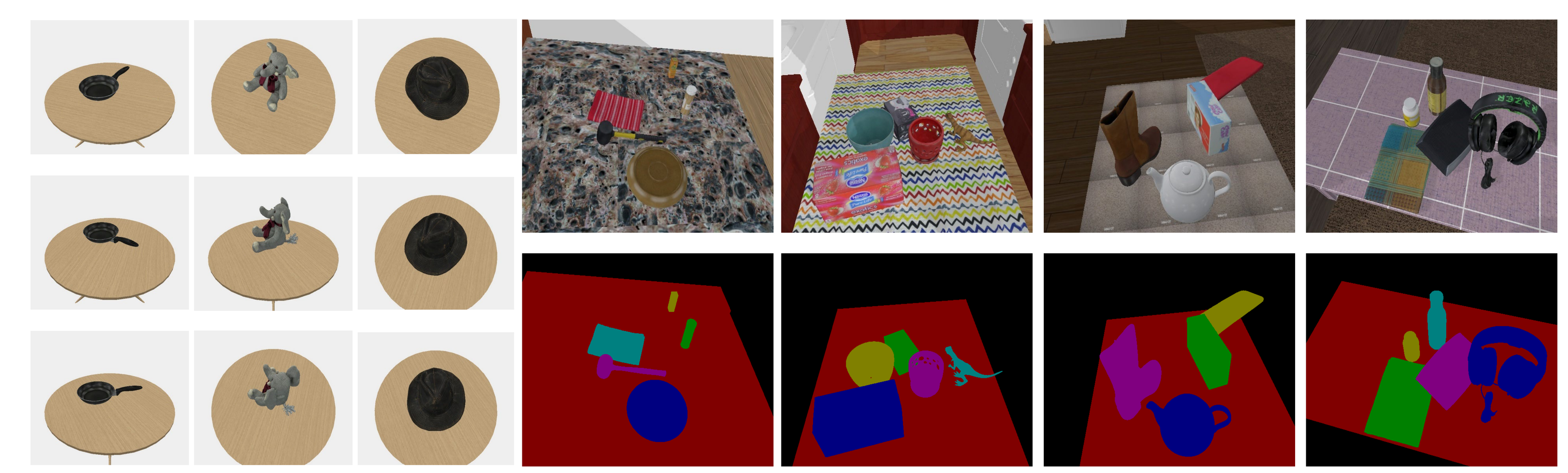
3. What is the object in these images made of? (list all materials of the object)
plastic, paper, metal

4. What can be the object in these images used for? (list all function of the object)
placing the tape, stick, wrap, separate

5. What is the color of the object in these images? (list all colors of the object)
black, transparent, white, yellow

Our Amazon Mechanical Turk questionnaire for object annotation.

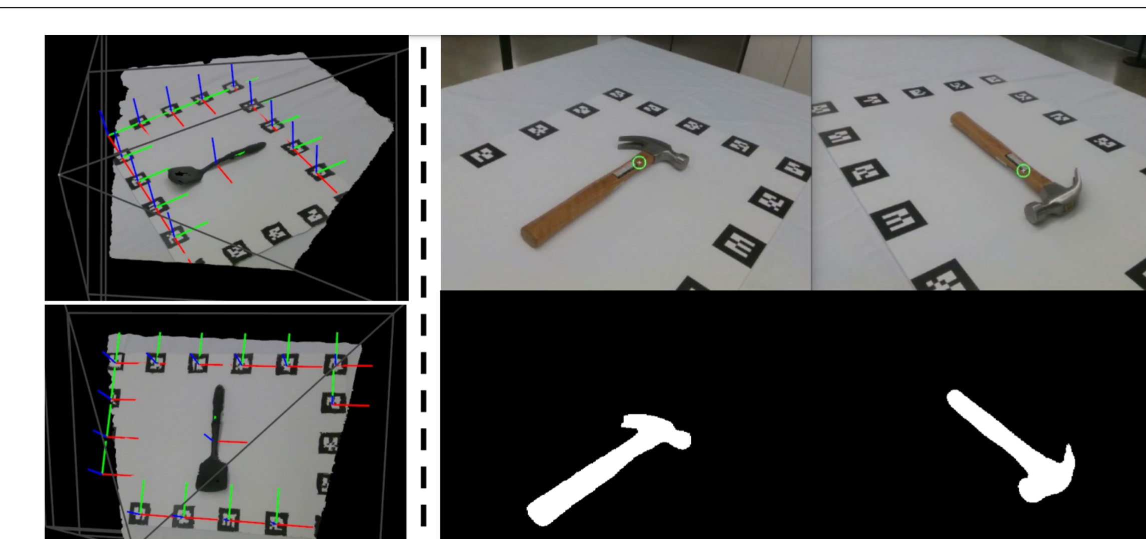
Dataset Construction: 3. Synthetic Data Generation



(a) (b)

(a) Synthetic objects with clean background. (b) Synthetic objects in cluttered scenes.

Dataset Construction: 4. Pose and Segmentation Annotations



(a) (b)

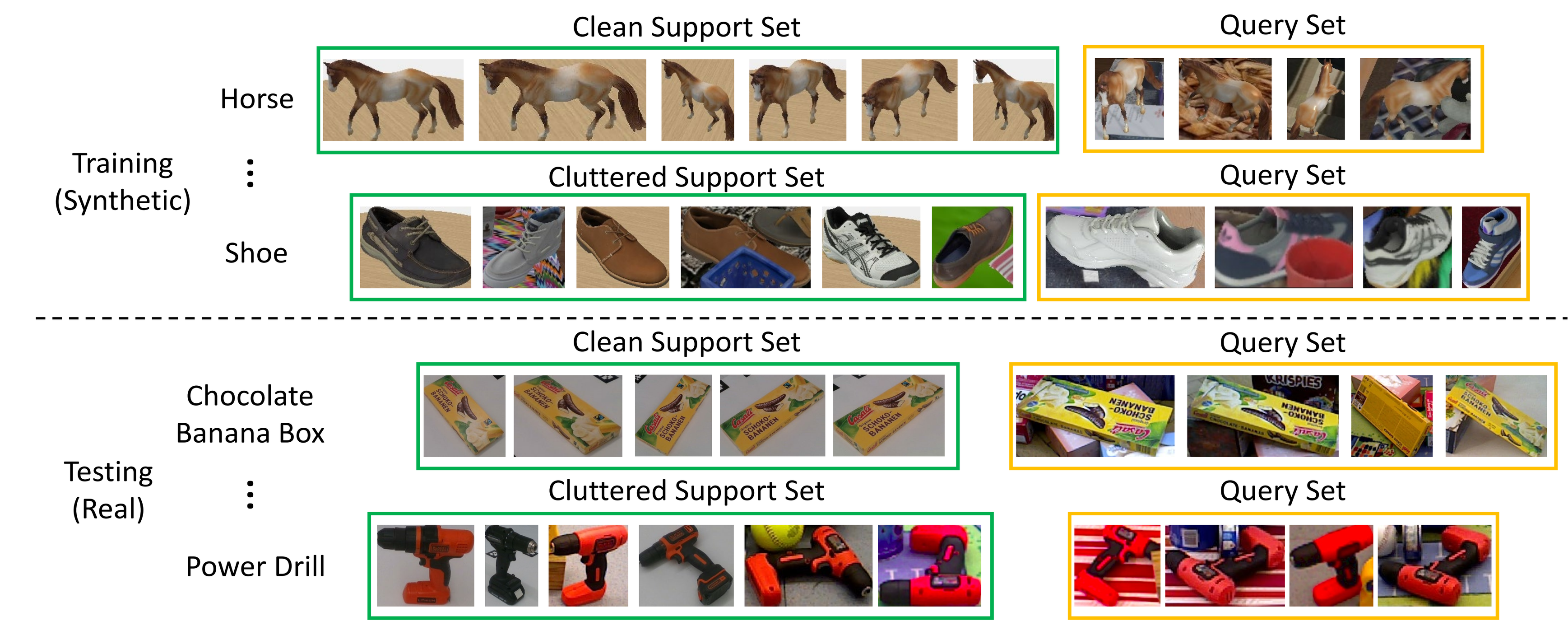
(a) Object poses from AR tags (b) Pixel correspondences using computed object poses and the segmentation masks of the objects.

Acknowledgments

This work was supported in part by the **DARPA Perceptually-enabled Task Guidance (PTG)** Program under contract number **HR00112220005**.

Scan me!

Experiments: Benchmarking using Meta-Dataset



Joint Object Segmentation and Few-Shot Classification

Method	OCID (Real)					
	Use GT segmentation (#classes, #objects)			Use custom segmentation (#classes, #objects)		
	All (52, 2300)	Unseen (41, 1598)	Seen (11, 702)	All (52, 2300)	Unseen (41, 1598)	Seen (11, 702)
	Clean S	Clean S	Clean S	Clean S	Clean S	Clean S
Training setting: clean support set with pre-training (top-1, top-5)						
k-NN	14.65, 25.22	15.33, 24.41	41.03, 72.65	12.70, 23.22	13.70, 22.59	36.75, 67.95
Finetune	22.26, 50.17	**26.41**, 58.20	31.62, 80.34	21.30, 48.57	**24.34**, 53.94	35.47, 67.38
ProtoNet	**25.17**, **57.30**	25.22, **58.45**	51.99, **94.73**	**22.96**, **51.96**	22.65, **54.32**	49.86, **87.75**
MatchingNet	17.39, 48.35	14.64, 50.06	51.85, 90.31	15.78, 45.13	13.08, 46.93	49.15, 84.47
fo-MAML	11.43, 31.48	11.58, 34.73	36.89, 69.94	10.91, 29.17	10.01, 32.35	31.77, 63.68
fo-Proto-MAML	14.35, 28.96	5.63, 40.61	45.58, 71.51	13.39, 26.96	5.51, 37.73	41.74, 67.24
CTX	17.48, 46.57	18.21, 49.81	51.85, 87.75	15.70, 43.83	16.90, 46.31	47.86, 81.34
CTX+SimCLR	18.57, 50.30	20.46, 51.06	**57.55**, 93.16	16.48, 46.17	17.71, 47.12	**52.14**, 85.75
Training setting: cluttered support set with pre-training (top-1, top-5)						
k-NN	13.70, 23.83	15.33, 24.28	47.72, 72.79	13.26, 23.22	14.14, 22.90	44.73, 68.66
Finetune	22.17, 53.35	24.34, 55.63	31.91, 71.51	18.26, 44.22	20.65, 52.00	36.04, 69.52
ProtoNet	21.35, 50.57	22.34, 51.31	51.99, 90.46	18.61, 47.22	18.21, 48.12	45.44, 85.33
MatchingNet	17.52, 50.96	17.77, 52.32	49.43, 88.18	16.52, 46.52	15.58, 48.81	43.45, 82.76
fo-MAML	16.48, 38.52	13.70, 39.49	37.46, 77.07	15.35, 35.04	11.08, 34.36	40.31, 69.94
fo-Proto-MAML	11.04, 28.70	4.01, 38.67	43.73, 72.65	9.91, 26.35	3.57, 35.79	40.46, 68.09
CTX	19.00, 45.48	17.71, 44.74	51.85, 88.75	17.13, 42.22	16.08, 42.12	47.15, 83.19
CTX+SimCLR	**24.61**, **62.39**	**25.16**, **63.52**	**65.81**, **96.30**	**22.17**, **57.43**	**23.28**, **57.57**	**59.12**, **88.32**
Using pre-trained CLIP models						
Few-shot Tip-Adapter ViT-L/14-Finetune	**60.17**, 83.04	**59.64**, 85.17	85.75, **99.00**	**54.87**, **78.91**	**56.07**, 80.29	79.20, 91.88
Few-shot Tip-Adapter ViT-L/14	56.78, 83.22	55.38, 84.86	**86.89**, 98.58	52.35, 76.26	51.69, 79.04	**80.06**, **92.45**
Zero-shot CLIP ViT-L/14	54.57, **84.74**	55.94, **87.92**	83.62, 98.58	50.43, 78.52	52.07, **81.54**	75.07, 92.17
Zero-shot CLIP ViT-B/32	41.87, 75.26	41.30, 77.91	78.06, 97.58	39.83, 69.43	39.17, 72.09	70.66, 90.88
Zero-shot CLIP ViT-B/16	40.70, 73.96	40.24, 76.03	76.50, 95.73	39.35, 68.83	38.61, 70.15	70.66, 88.89
Zero-shot CLIP RN50x64	42.96, 75.83	43.62, 77.41	76.64, 96.01	40.04, 70.87	41.74, 72.22	69.94, 90.46
Zero-shot CLIP RN50x16	38.52, 73.04	40.11, 75.72	79.49, 96.30	35.65, 67.30	37.30, 69.77	70.94, 89.74
Zero-shot CLIP RN50x4	35.96, 68.52	34.42, 70.03	73.93, 95.73	34.00, 63.78	32.48, 65.46	67.95, 88.60
Zero-shot CLIP ResNet-101	32.96, 68.30	32.67, 69.52	77.49, 96.87	31.09, 63.87	31.85, 65.96	69.66, 89.74
Zero-shot CLIP ResNet-50	25.91, 58.43	29.04, 64.39	61.40, 93.16	24.70, 55.61	28.04, 61.20	57.69, 86.47

Qualitative Results in the Real World