7th Conference on Robot Learning (2023), Atlanta, USA

STOW: Discrete-Frame Segmentation and Tracking of Unseen Objects for Warehouse Picking Robots

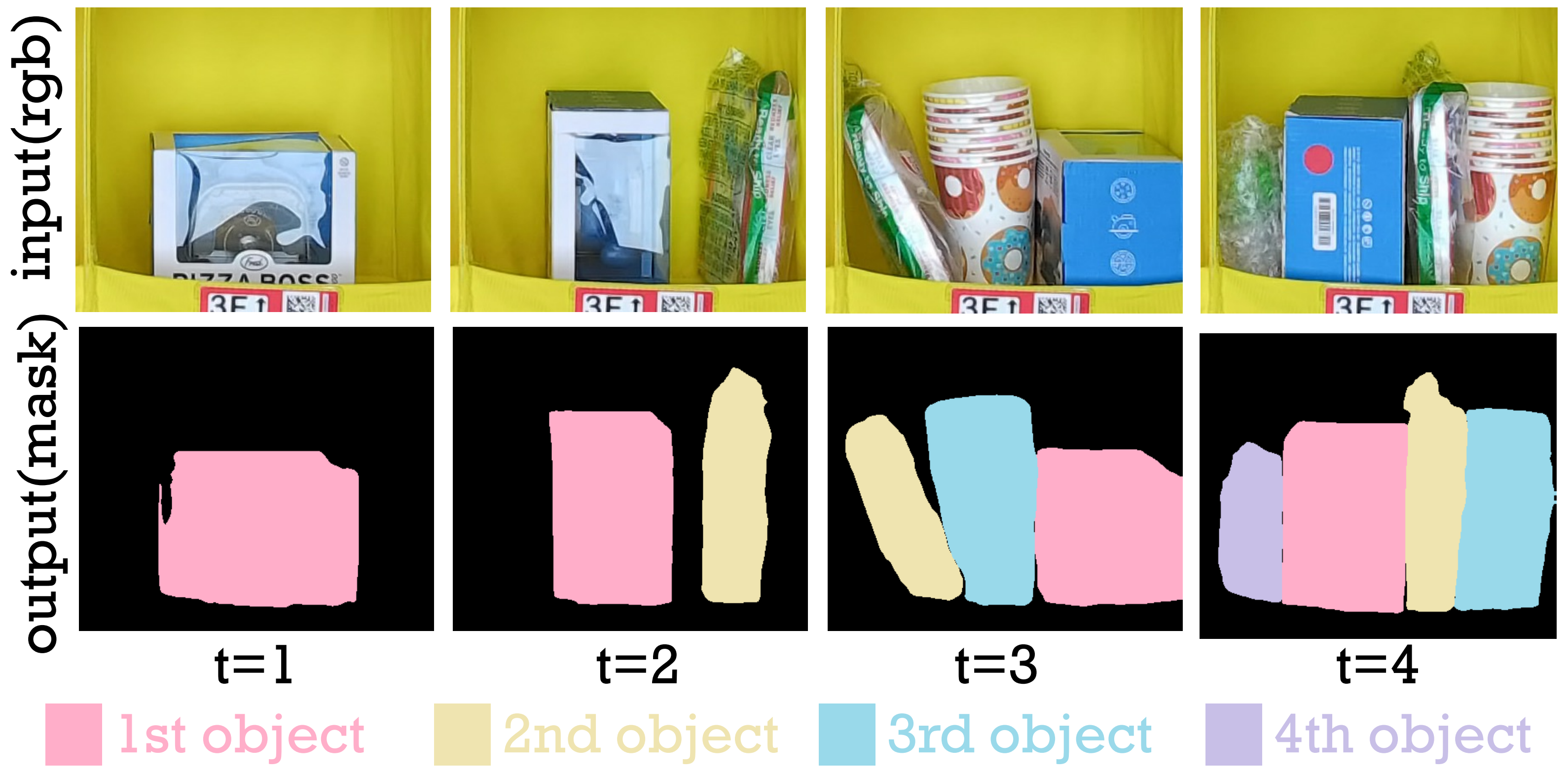
Yi Li[1], Muru Zhang[1], Markus Grotz[1], Kaichun Mo[2], Dieter Fox[1,2]
University of Washington, NVIDIA

CODE, DATA AND MORE

SCAN ME

Task Formulation

- ◆ A cluttered shelf contains diverse objects
- ◆ objects may be rearranged
- ◆ camera may be occluded
- ◆ The goal is to pick an object based on its given order index in the bin.



$$\mathcal{I} = \{I_1, I_2, \cdots, \boxed{I_T} \mid I_t \in \mathbb{R}^{H\times W\times C_I}\}$$

T: num of frames
K_I: num of objects

$$\mathcal{M}_t = \{M_t^1, M_t^2, \cdots, M_t^{K_I} \mid M_t^i \in \{0,1\}^{H\times W}, i = 1, 2, \cdots, K_I\}$$

Challenges

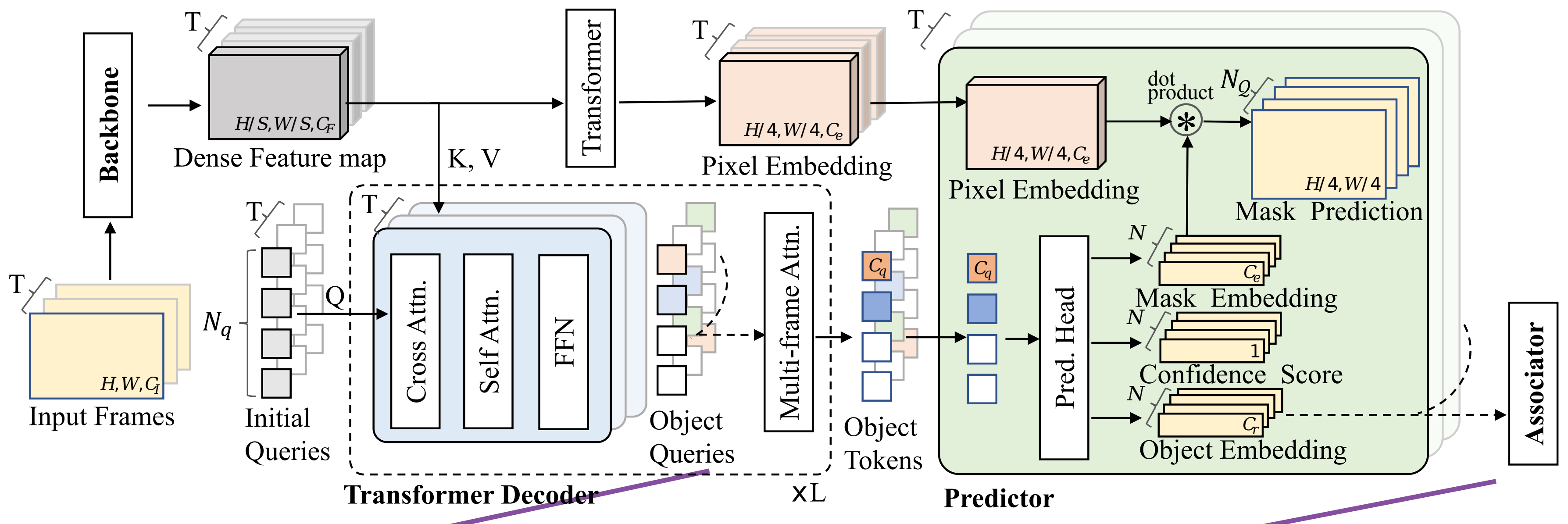
Unseen Objects

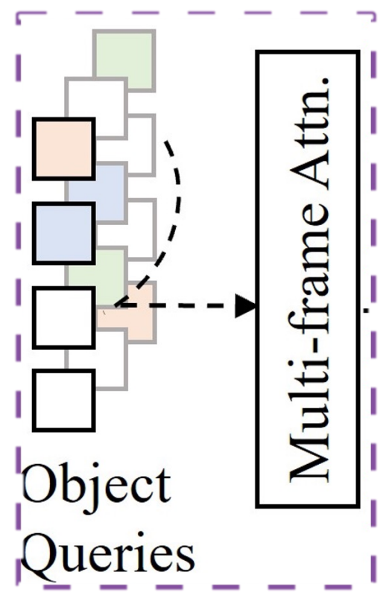
- Vast number of categories
- Limited synthetic models
- Ambiguous border
- Transparent parts
- Plastic wrapper

Discrete frames

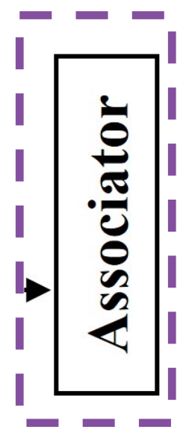
- Caused by heavy occlusion
- Drastic appearance change
- Large movement between frames
- Little information from context

Method





- Multi-frame attention
- On object queries from different frames
- Inside decoder block
- Efficient inter-frame reasoning

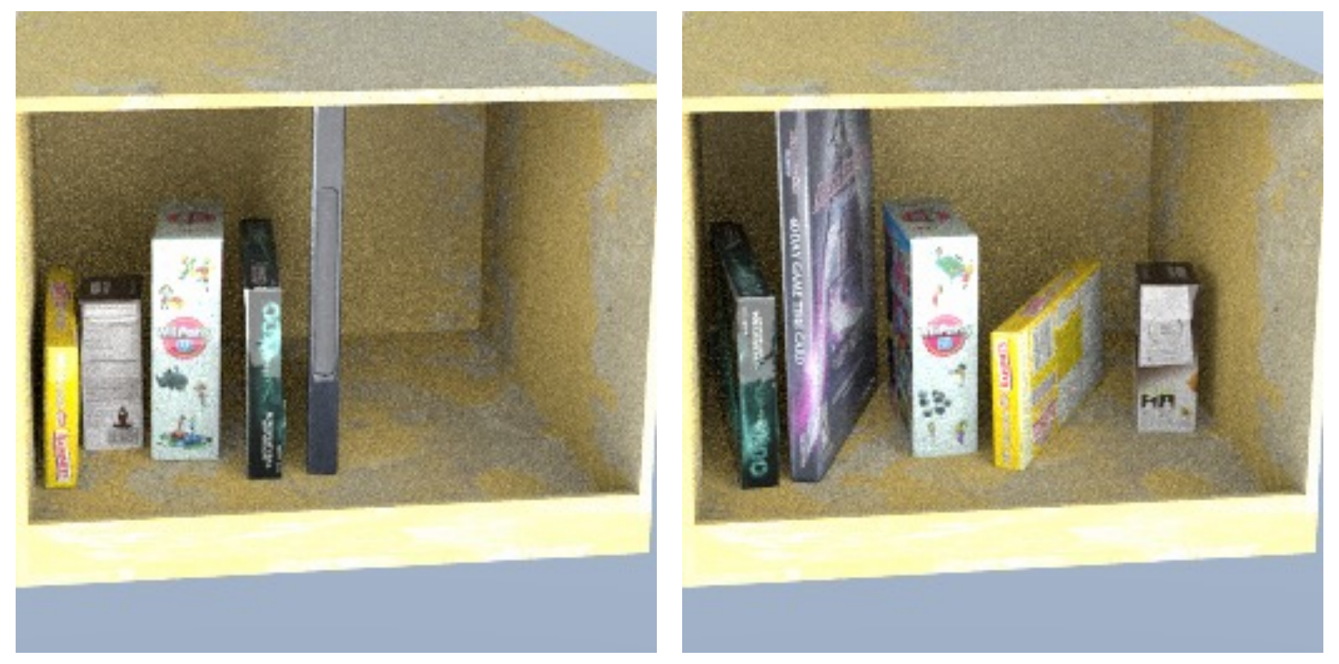


- Hungarian matching
- Embeddings for tracking
- From each frame
- Disentangles tracking from detection

Dataset

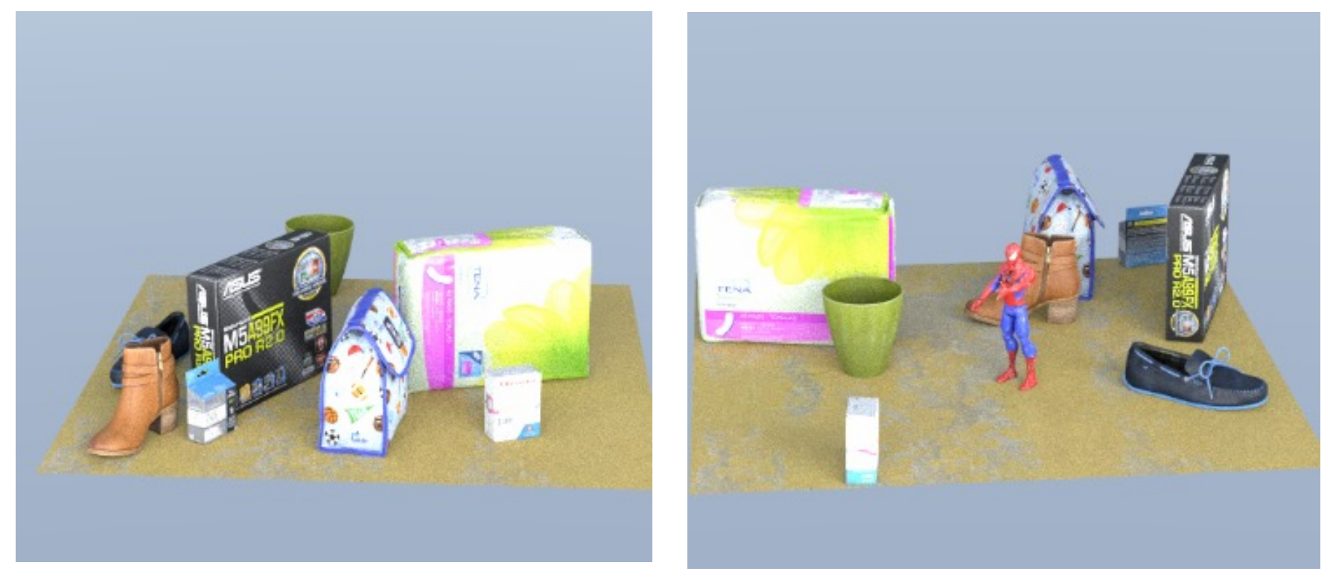
Synthetic Data (train & val)
GoogleScanned objects
~900 for train, ~100 for test

shelf



10k sequences
2 frames for each sequence

tabletop



2k sequences
15 frames for each sequence

Real Data (test)
~ 150 real objects
manual annotation

44 sequences of bins
220 images in total

20 sequences
280 images in total

Results

Method	Shelf		Tabletop	
	AP@all	AP@0.5	AP@all	AP@0.5
MinVIS	6.3	21.2	0.7	0.0
Mask2Former Video	35.0	66.1	27.7	56.7
VITA	42.7	70.1	26.6	55.0
STOW (Ours)	**55.6**	**81.3**	**49.7**	**75.4**

- Comparison with STOA VIS methods
- Train on synthetic data and test on real data
- All using RN50 backbone with same number of iteration

multi frame	shelf		tabletop	
	AP@all	AP@0.5	AP@all	AP@0.5
-	51.8	78.7	44.4	68.5
✓	**55.6**	**81.3**	**49.7**	**75.4**

- Ablation study on multi-frame attention layer
- Frame attention layer can boost performance by ~5%

method	synthetic		real	
	AP@all	AP@0.5	AP@all	AP@0.5
MinVIS	0.3	2.6	0.7	0.0
M2F-V	71.6	83.7	27.7	56.7
VITA	69.4	81.9	26.6	55.0
STOW (Ours)	**74.1**	**89.3**	**49.7**	**75.4**

- Better performance handling Sim2Real Gap
- Train on synthetic and test on synthetic and real

Real Robot Experiments

82 trials, involving >100 objects

Method	Success Rate
UCN+SIFT	40.2%
VITA	46.3%
STOW(Ours)	74.4%

Acknowledgement

This research is funded by the UW + Amazon Science Hub as part of the project titled, "Robotic Manipulation in Densely Packed Containers." We would like to thank Dr. Michael Wolf from Amazon for valuable discussions. We further would like to thank our students Sanjar Normuradov and Soofiyan Atar for helping running robot experiments.