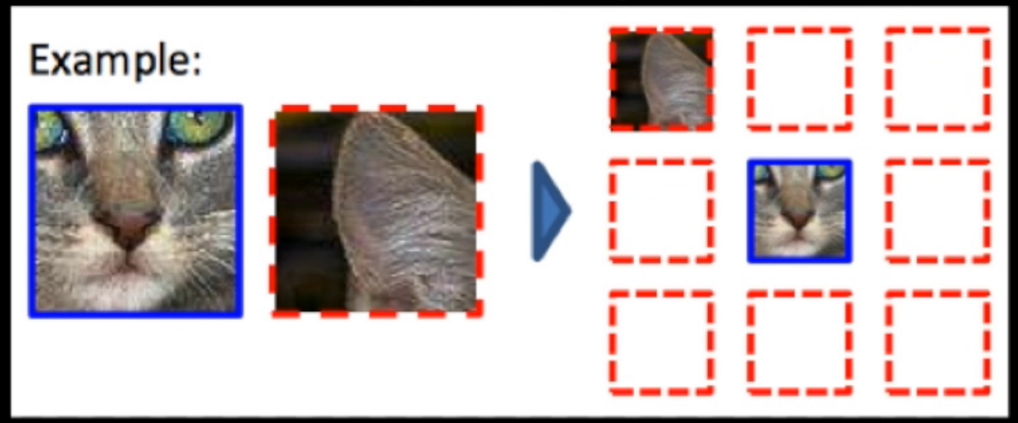
The Curious Robot

Learning Visual Representations via Physical Interactions

Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park & Abhinav Gupta

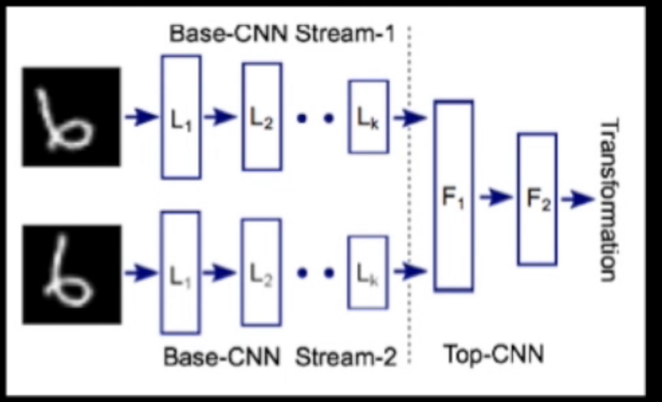
Unsupervised learning of visual features

Single Images



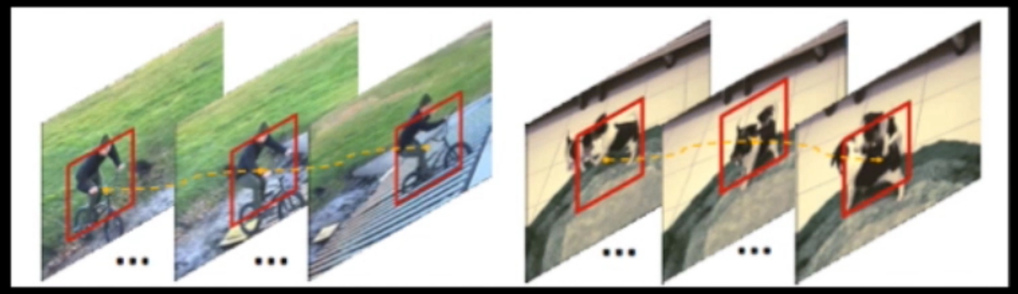
Doersch et al. 2015

Motion



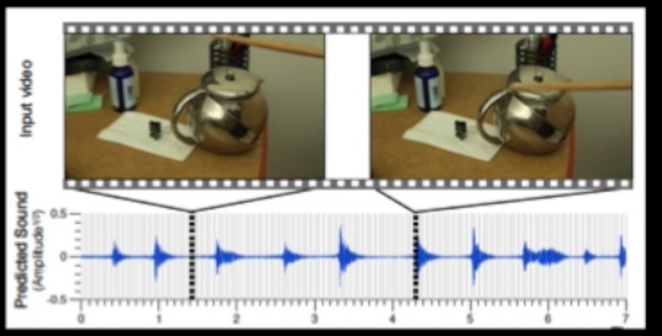
Agrawal et al. 2015

Videos



Wang et al. 2015

Sound



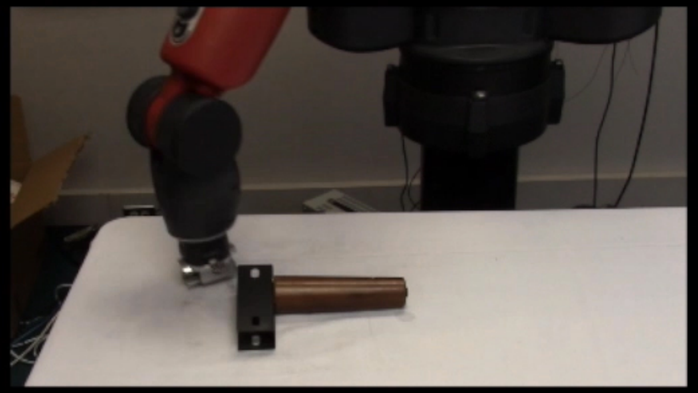
Owens et al. 2016

But how do **WE** learn these representations?

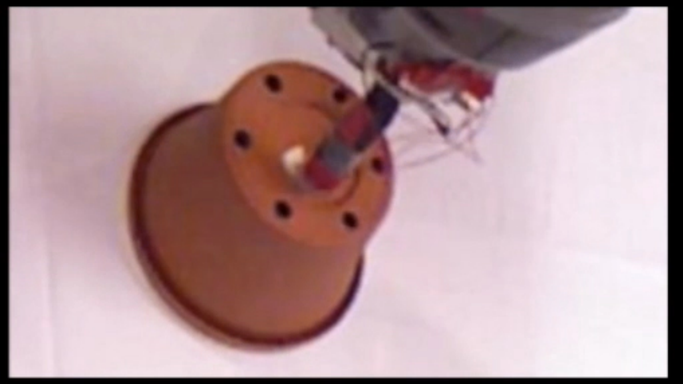
Using **robot** tasks to learn visual features

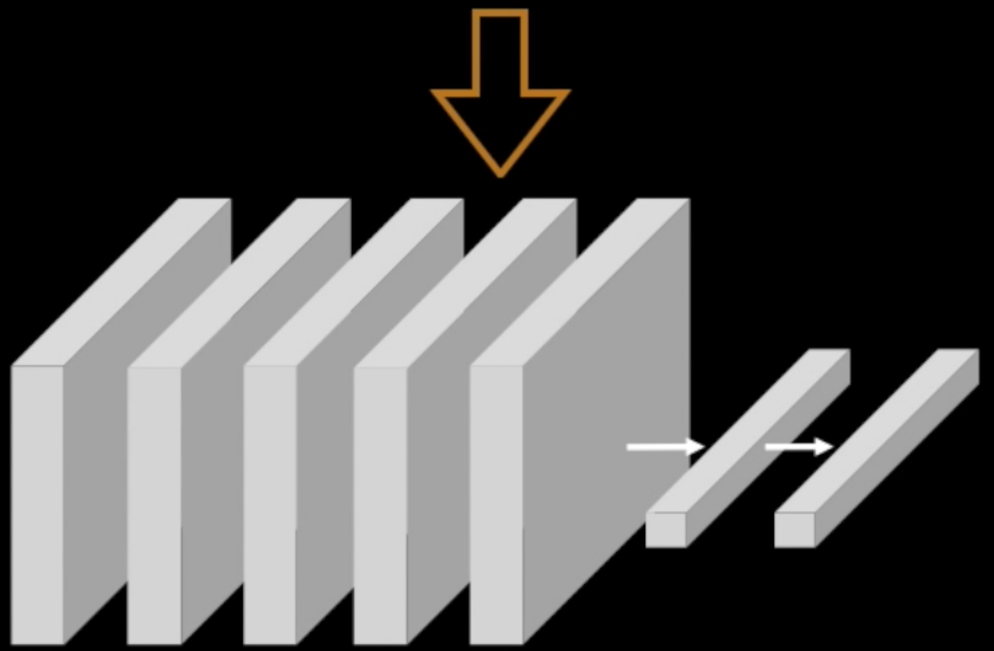
Grasping

Pushing



Poking





Using **robot** tasks to learn visual features

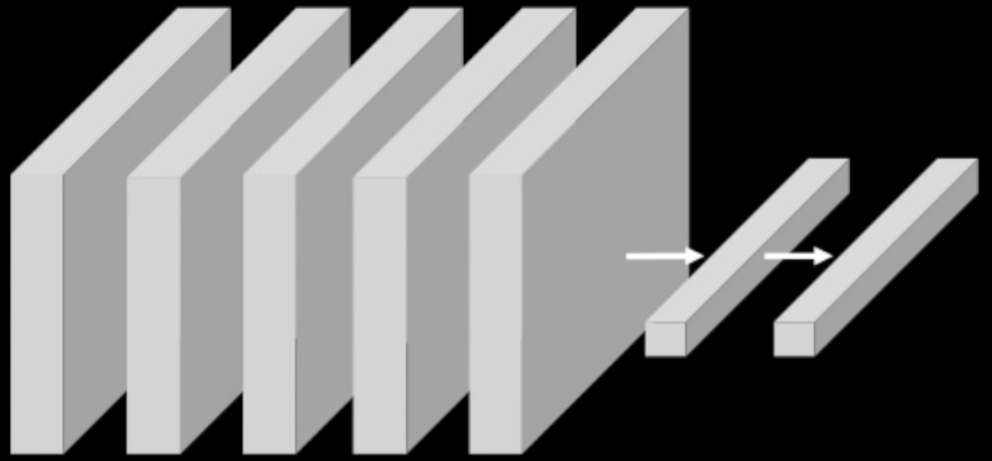
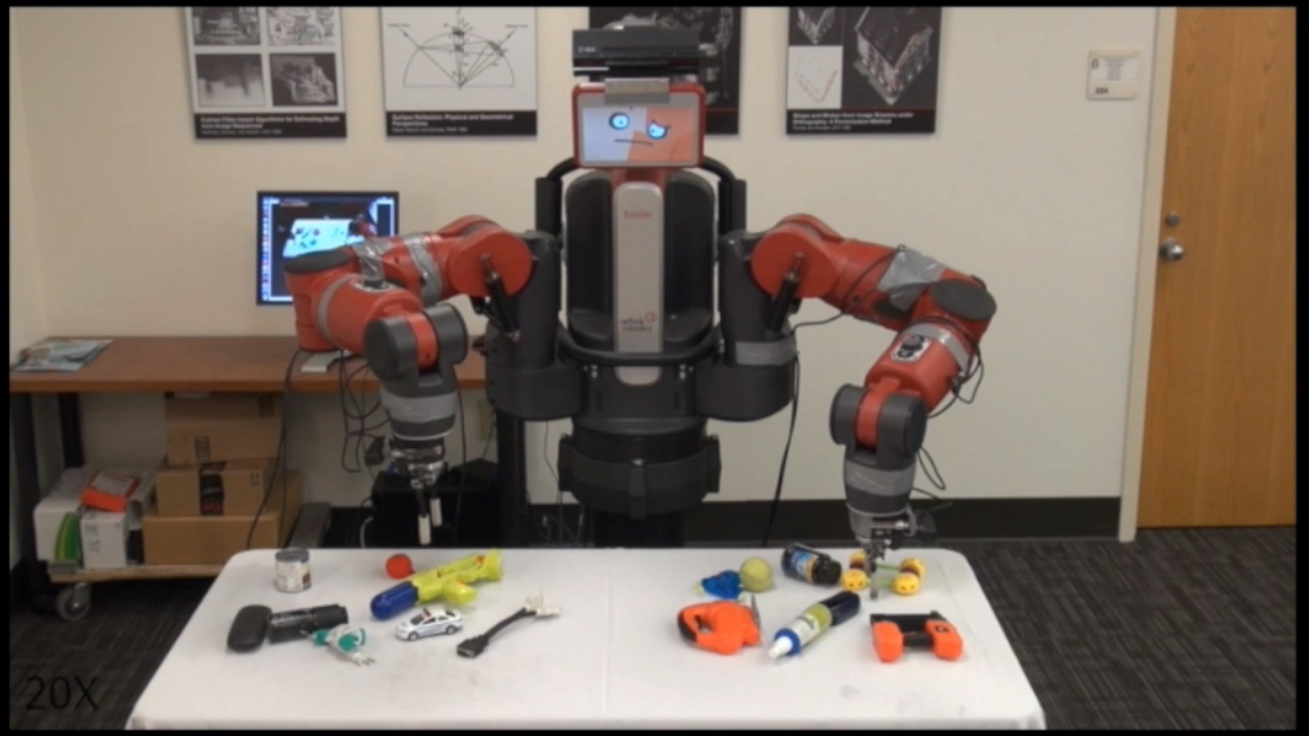


Image Classification

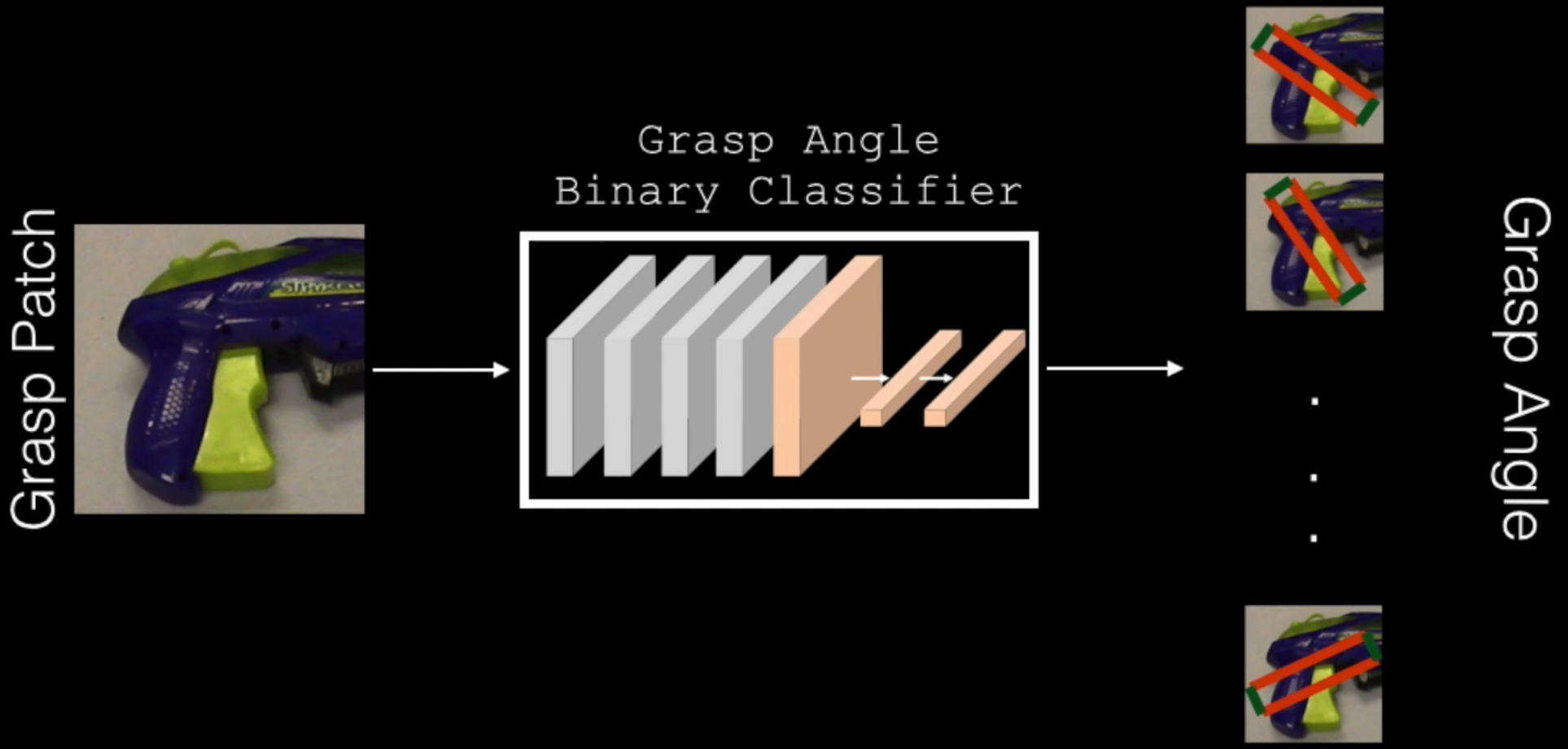
Image Retrieval

Physical Robot Tasks

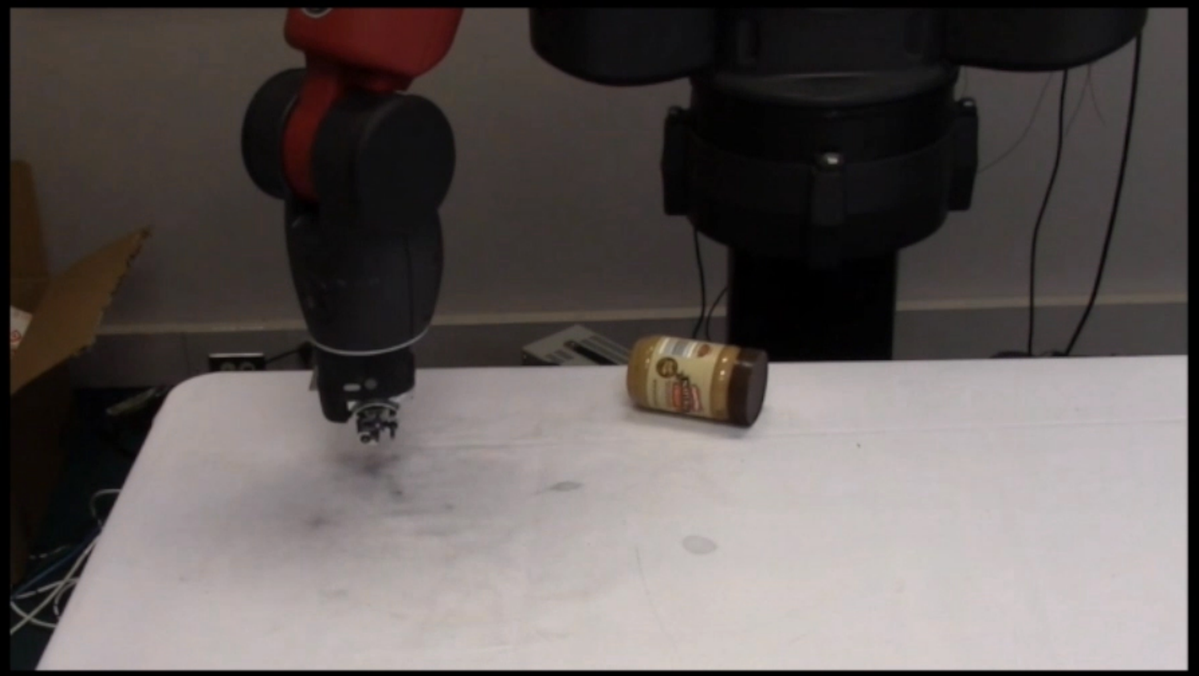
1. Planar Grasping

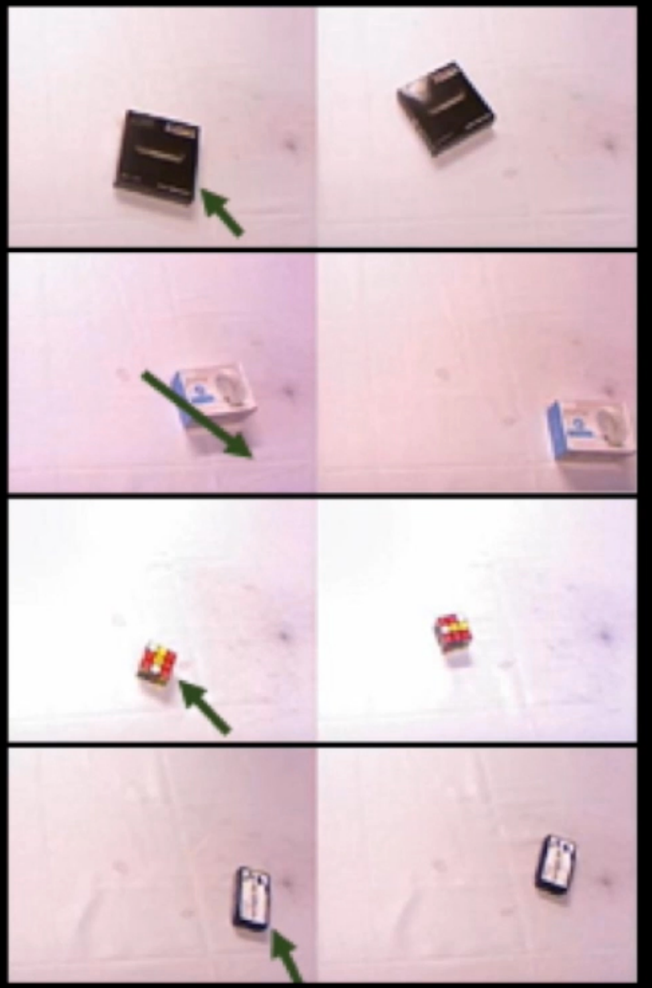


1. Planar Grasping Formulation

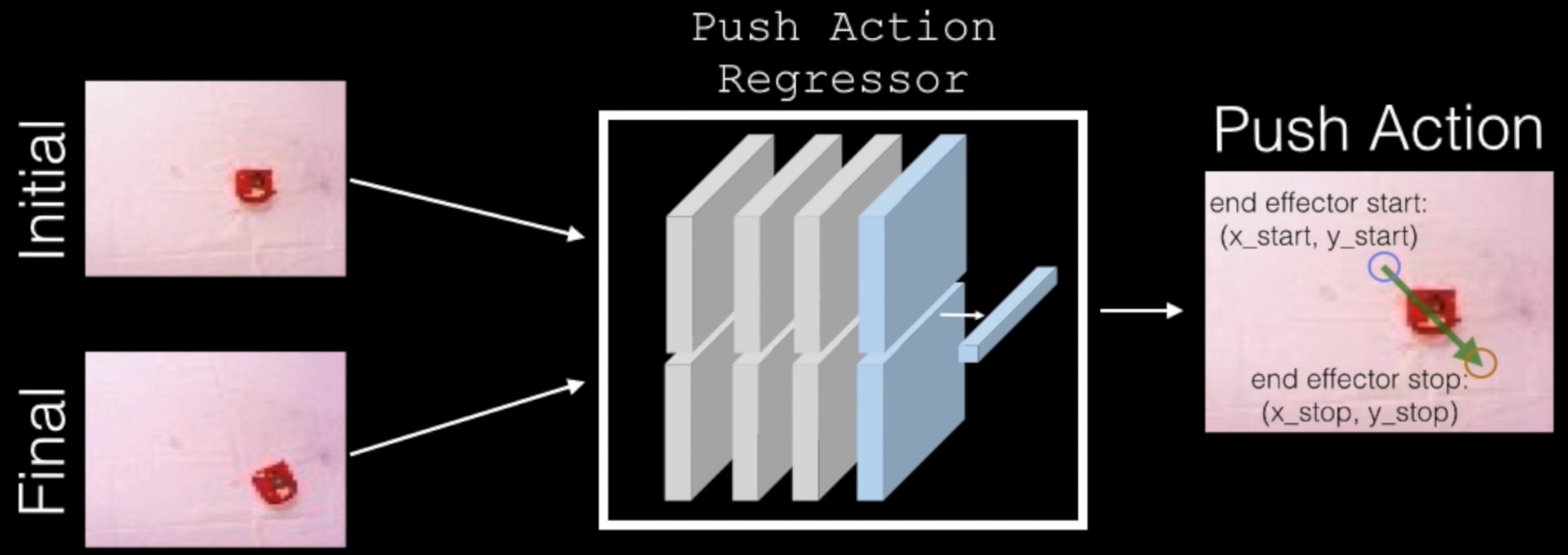


2. Planar Pushing



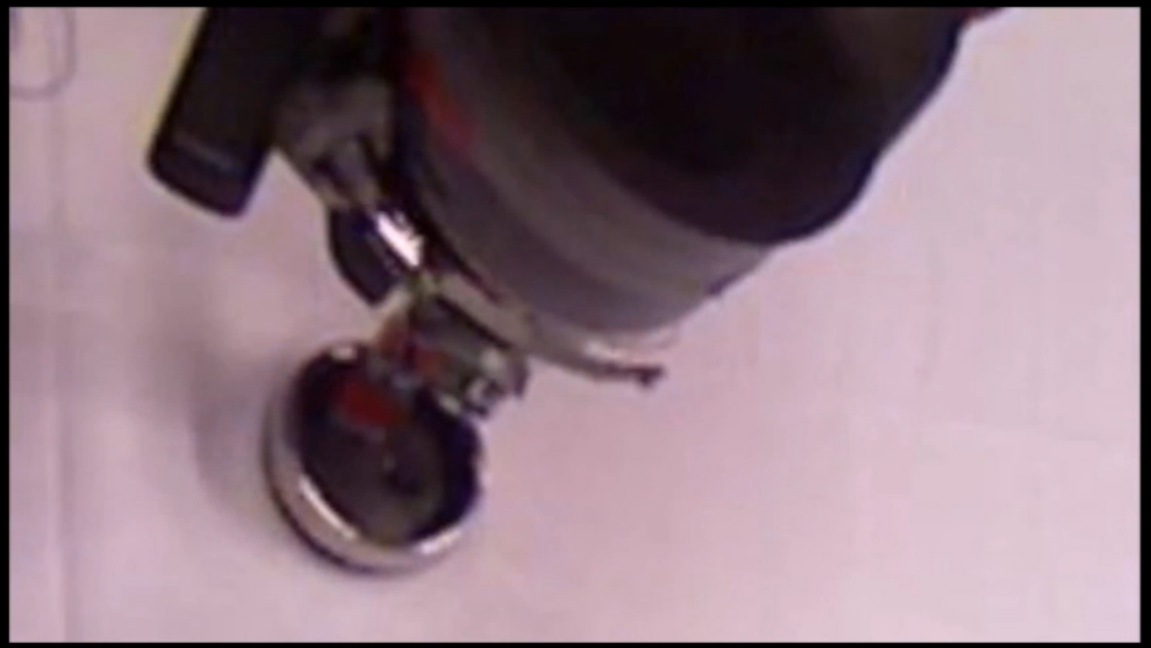


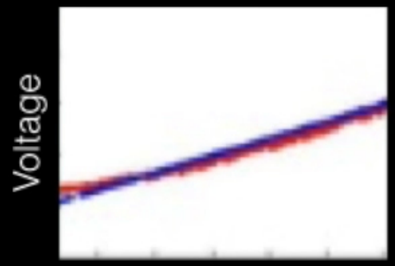
2. Planar Pushing Formulation

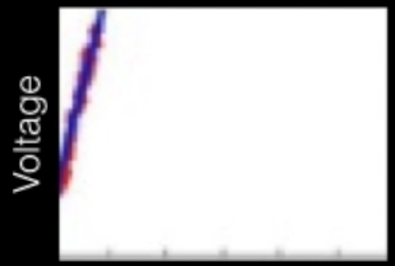


Push action parametrized by: (x_start, y_start, x_stop, y_stop, z)

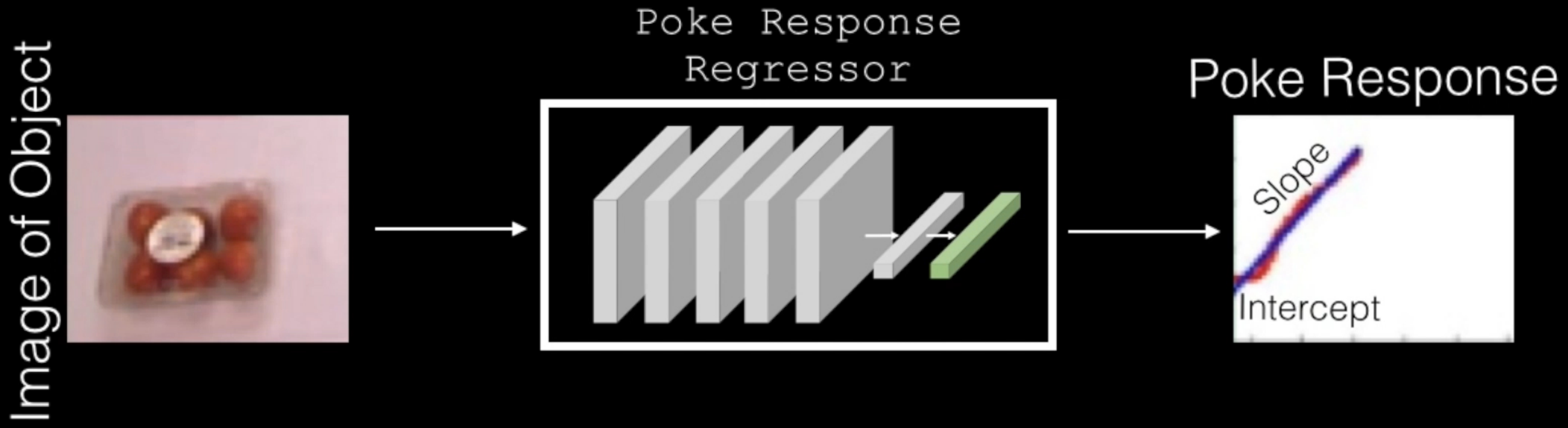
3. Poking: Tactile Sensing







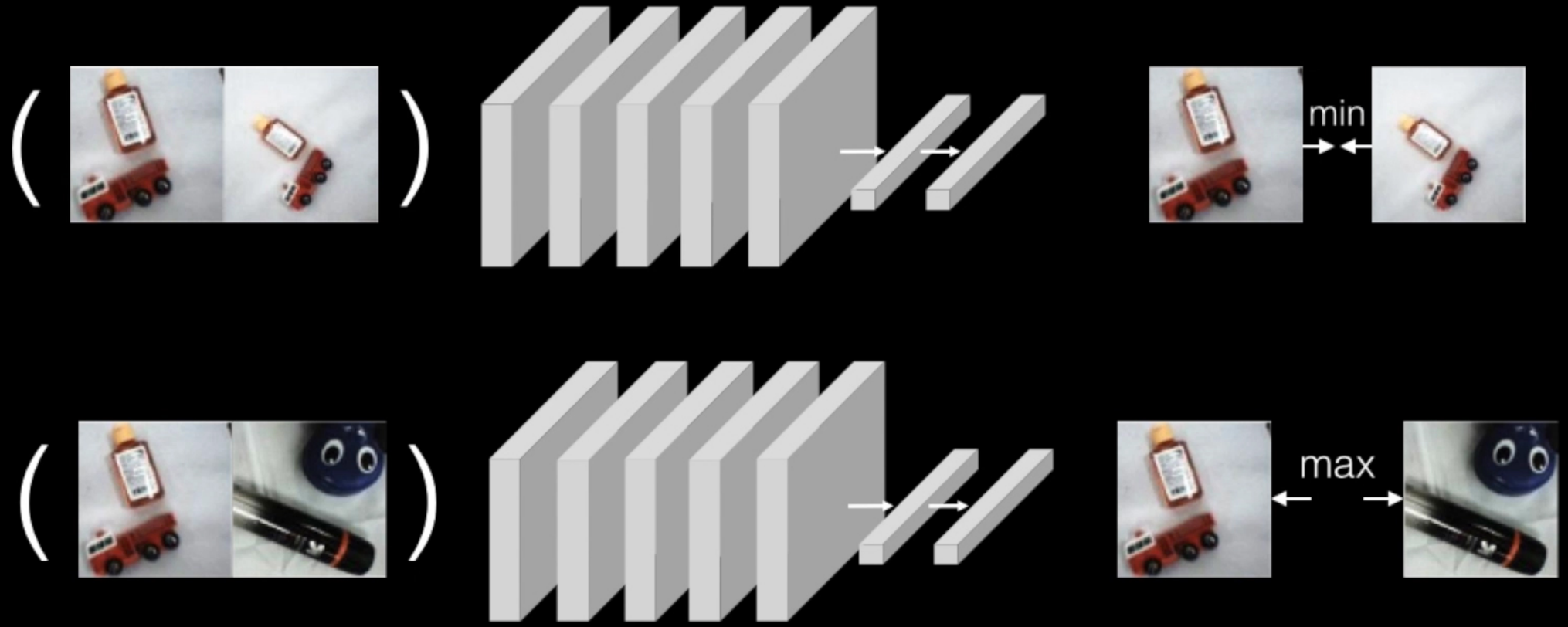
3. Poking: Tactile Sensing Formulation



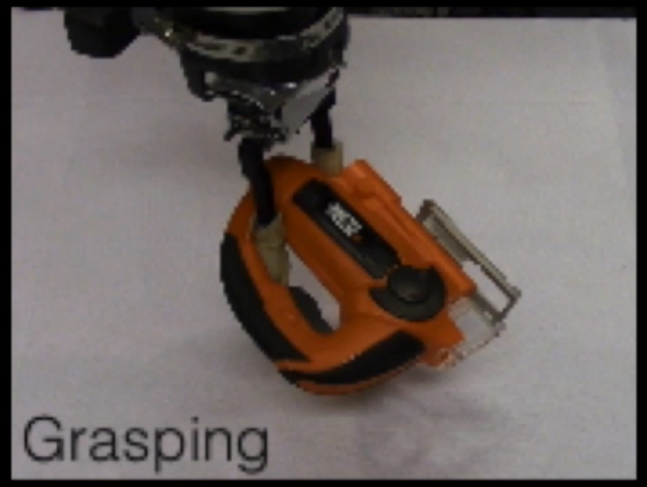
Poke response parametrized by: (slope, Intercept) of the tactile sensor response

4. Pose and Scale Invariance

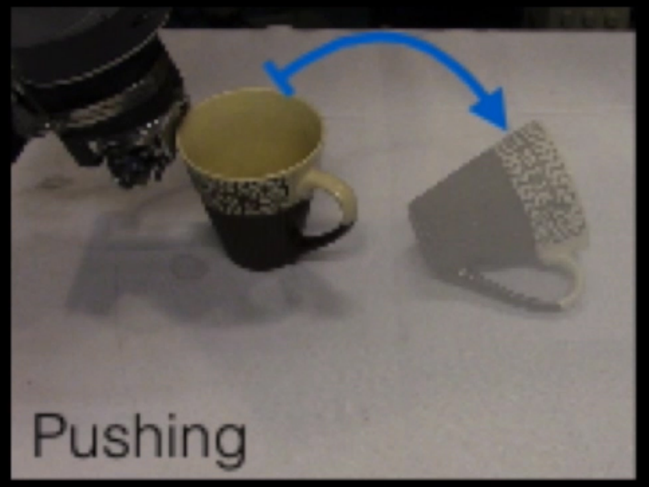
4. Pose and Scale Invariance Formulation



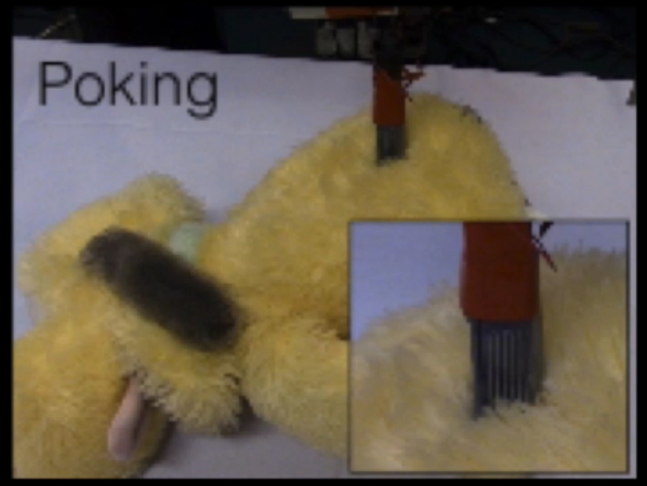
Multi-task Learning - Network Architecture



3K successful, 37K unsuccessful



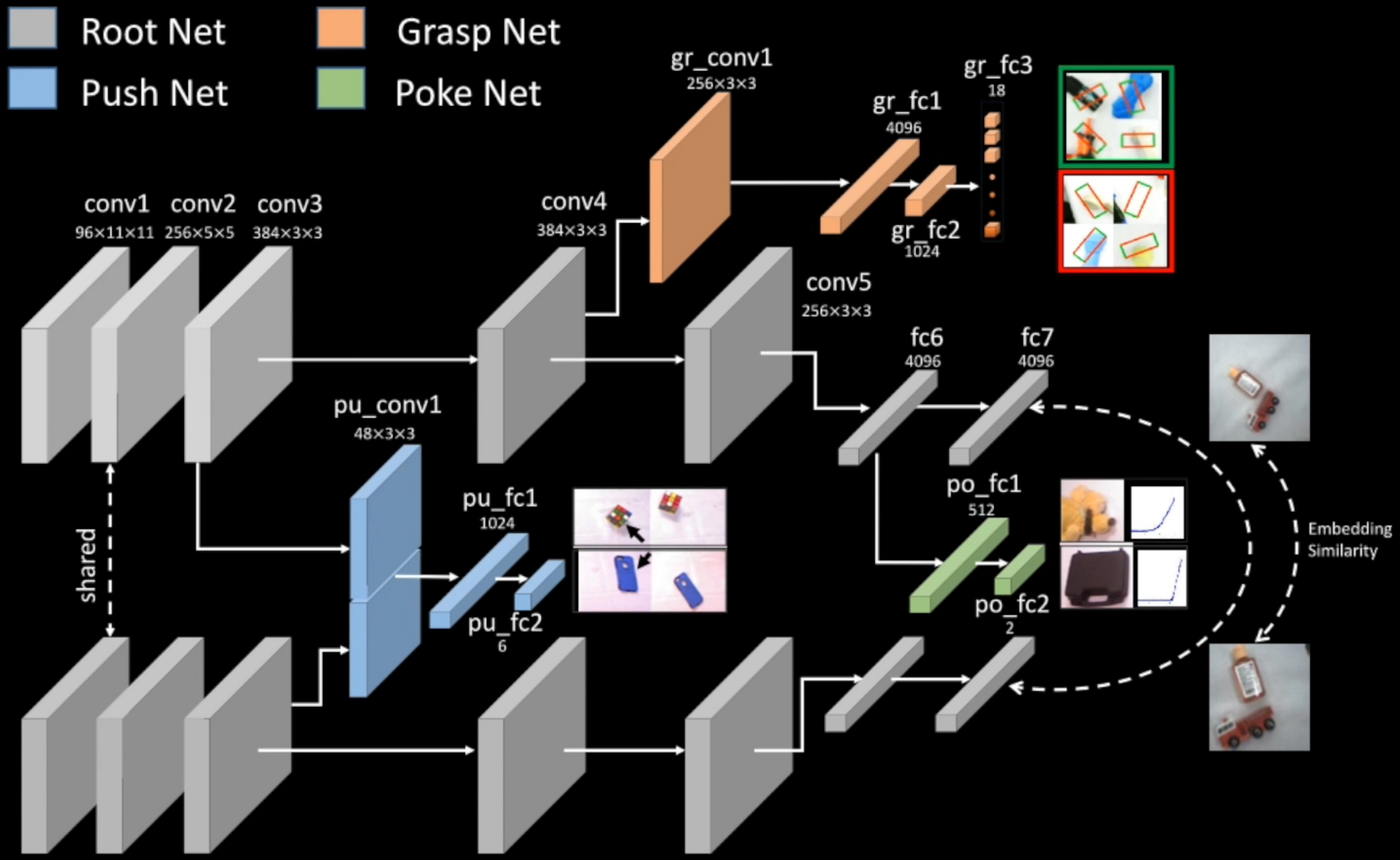
5K pushes



1K pokes

42K positive pairs, 42K negative pairs

Multi-task Learning - Network Architecture



Visual Representations?

UW RGB-D Category Classification

	Accuracy (%)	
Root network with random initialization	46.87	
Root network pretrained on robot tasks (ours)	69.37	**+22%**
AlexNet pretrained	82.03	

Ablation

	UW RGBD	
All (ours)	**69.3**	
Without Invariance	71.1	**+1.8%**
Without Grasp	63.2	**-6.1%**
Without Push	71.0	**+1.7%**
Without Poke	68.4	**-0.9%**

Instance level Image Retrieval

	recall@k=1	recall@k=5
randomNet	6.23%	21.93%
alexNet	68.60%	85.73%
ourNet	72.07% **+3.4%**	83.17%