More than a Feeling

Learning to Grasp and Regrasp using Vision and Touch

Presented by Lance Bassett

Paper by Calandra, Owens, Jayaraman, Lin, Yuan, Malik, Adelson, Levine

The Grasping Task

Task: Grasp and lift rigid, deformable, irregular objects

In humans: Active task involving reaching, placing fingers, balancing contact forces, and **adjusting**

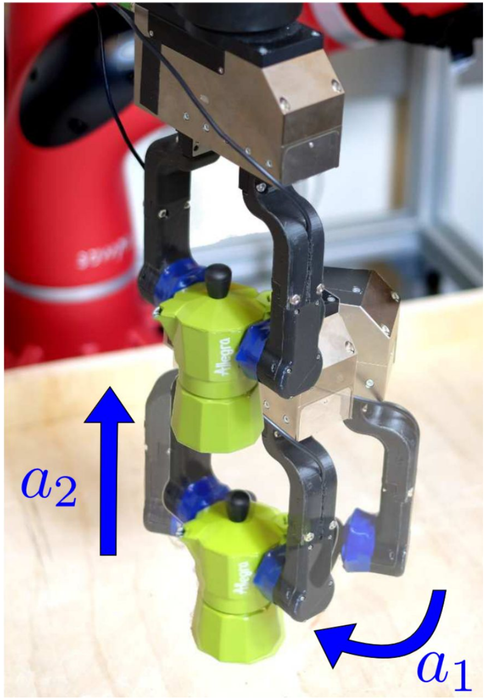
The Grasping Task

Task: Grasp rigid but irregular objects

In humans: Active task involving reaching, placing fingers, balancing contact forces

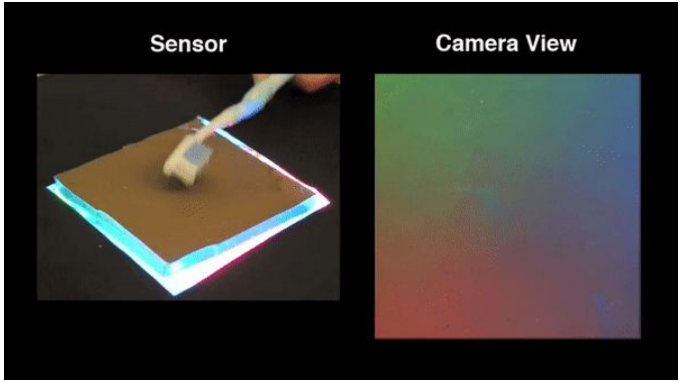
Previously	This Paper
Vision only →	**Visuo-tactile**
Feature engineering/Analytic Modeling →	**Data-driven**
Single grasp prediction →	**Iterative regrasping**

Hardware Components



Robot Grasping Arm

GelSight Feedback (Tactile)

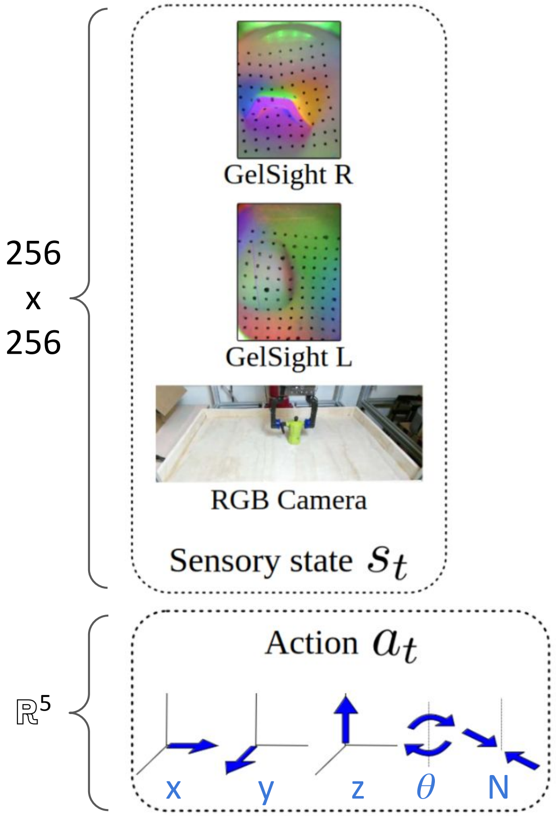


+

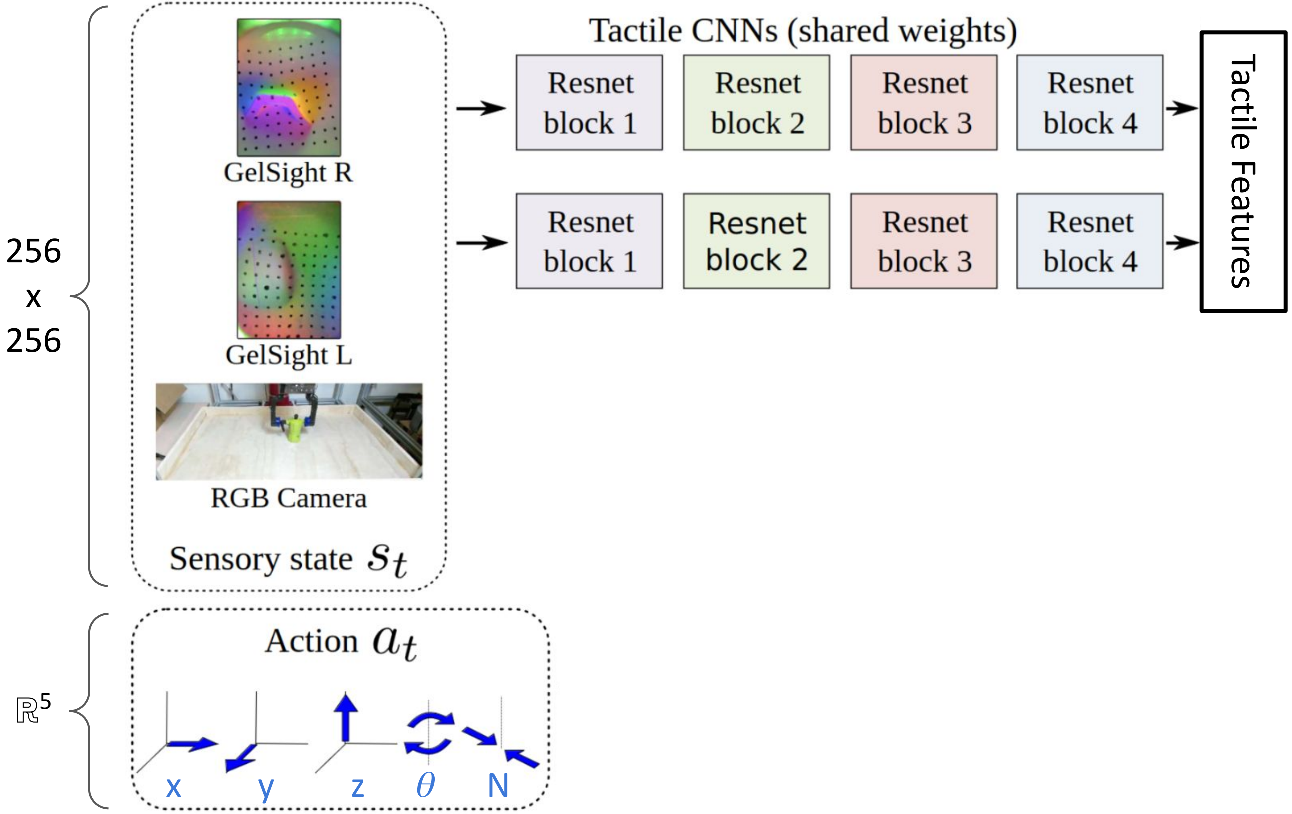
=

How should I adjust grip in order to successfully lift this object?

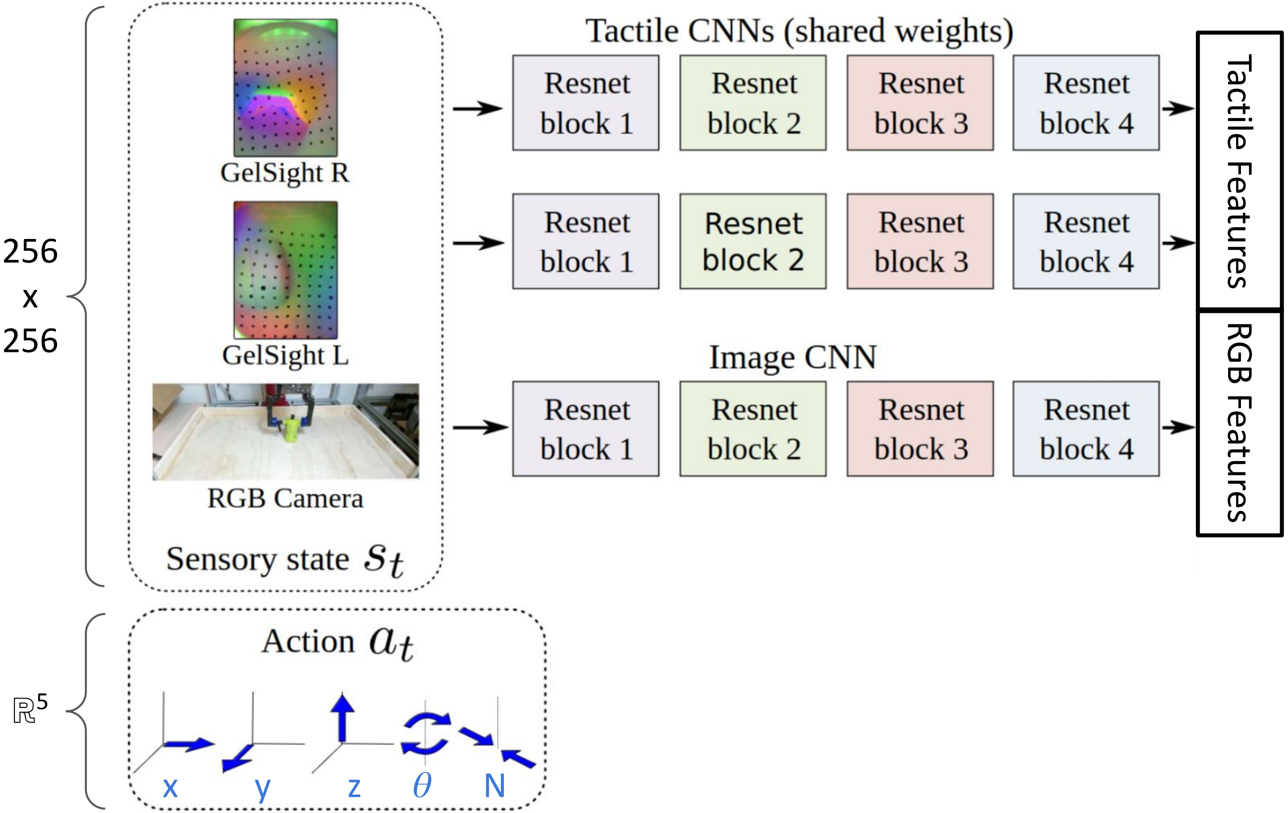
Network Design: Action-conditioned model *f*(*s*,*a*)



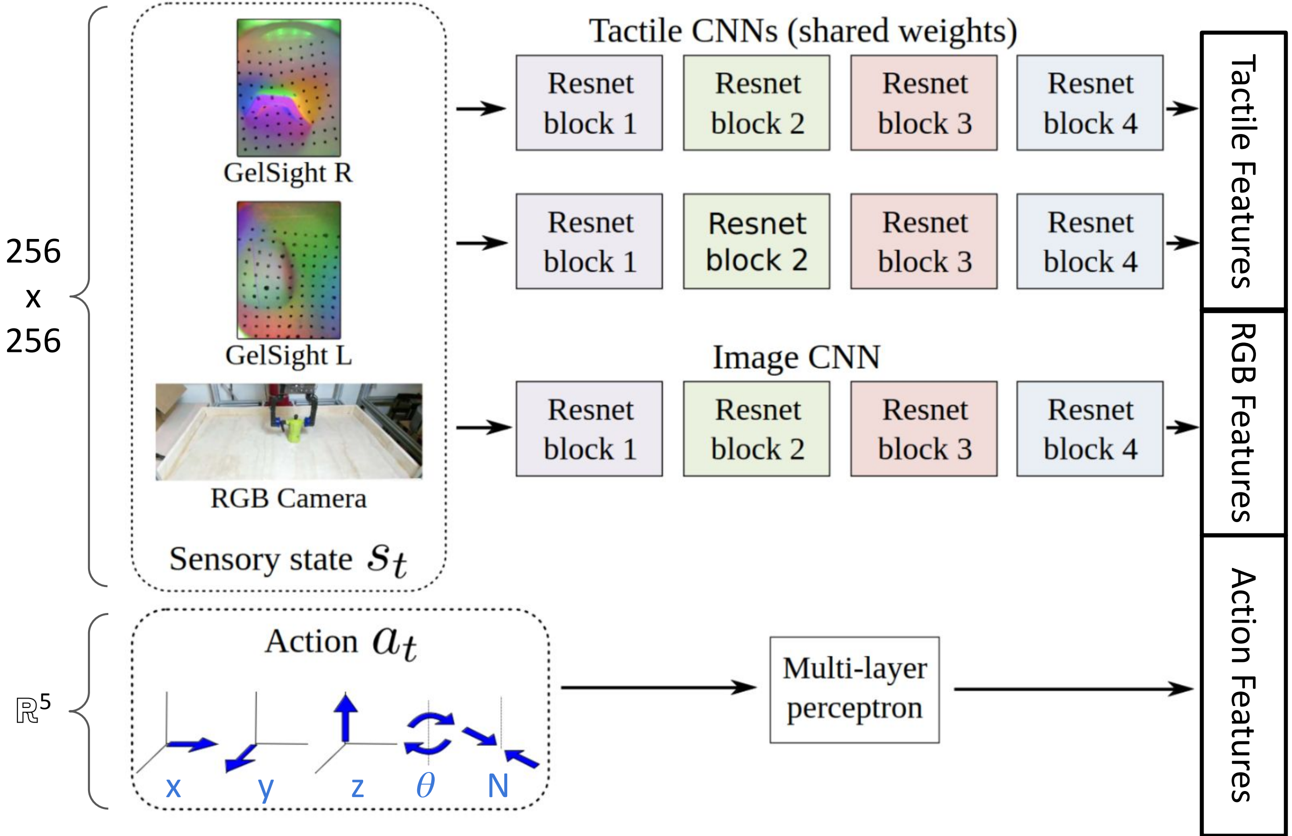
Network Design: Action-conditioned model ***f*(*s*,*a*)**



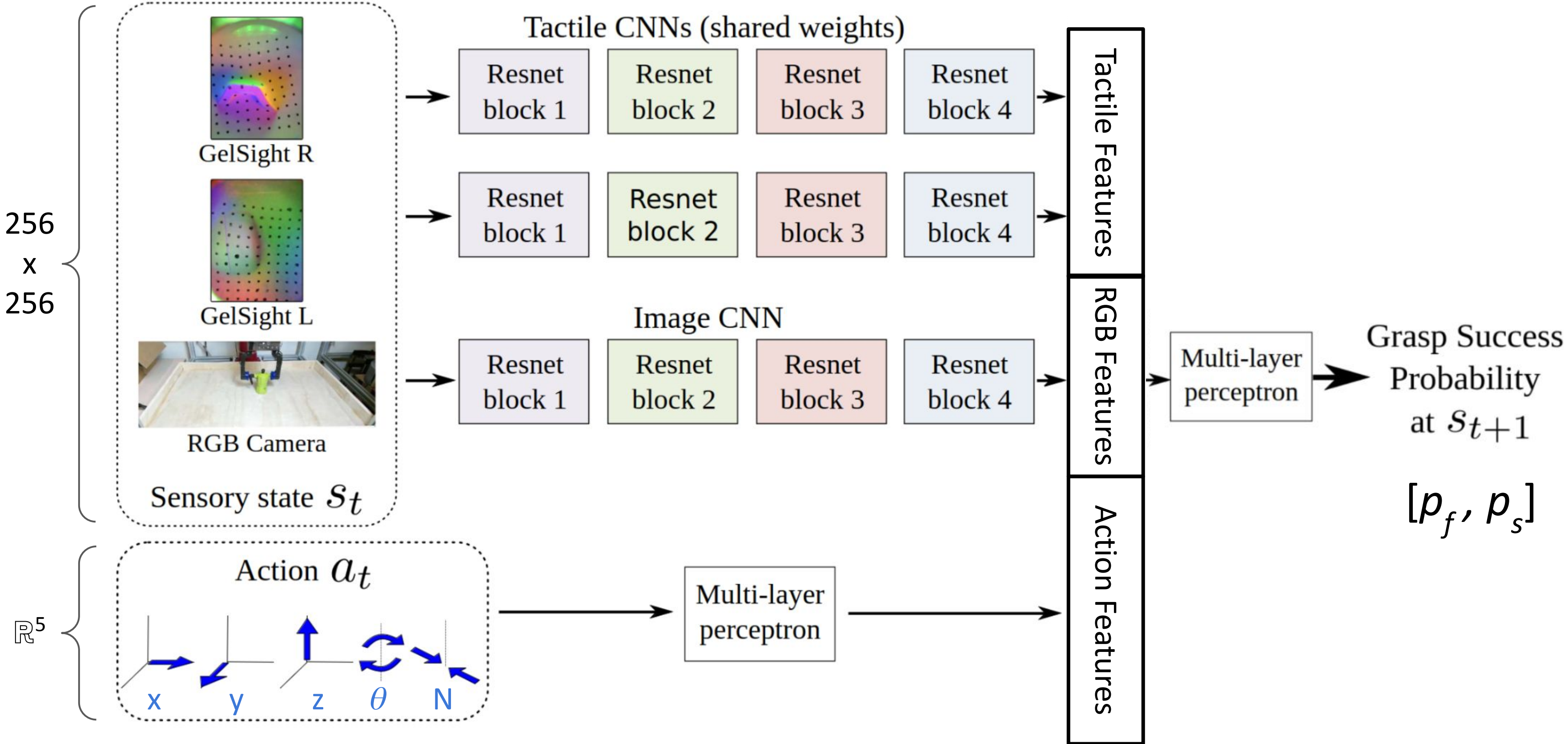
Network Design: Action-conditioned model ***f*(*s*,*a*)**



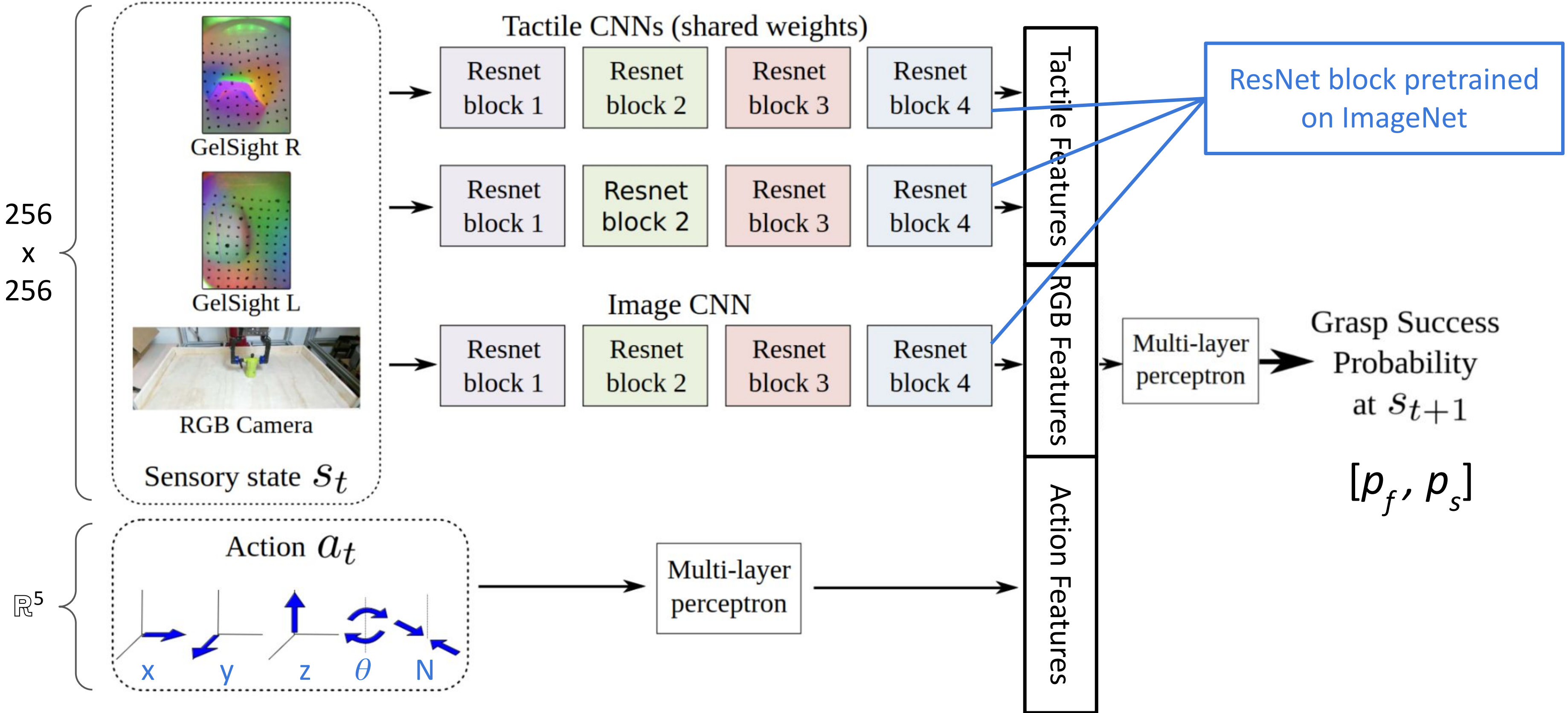
Network Design: Action-conditioned model ***f(s,a)***



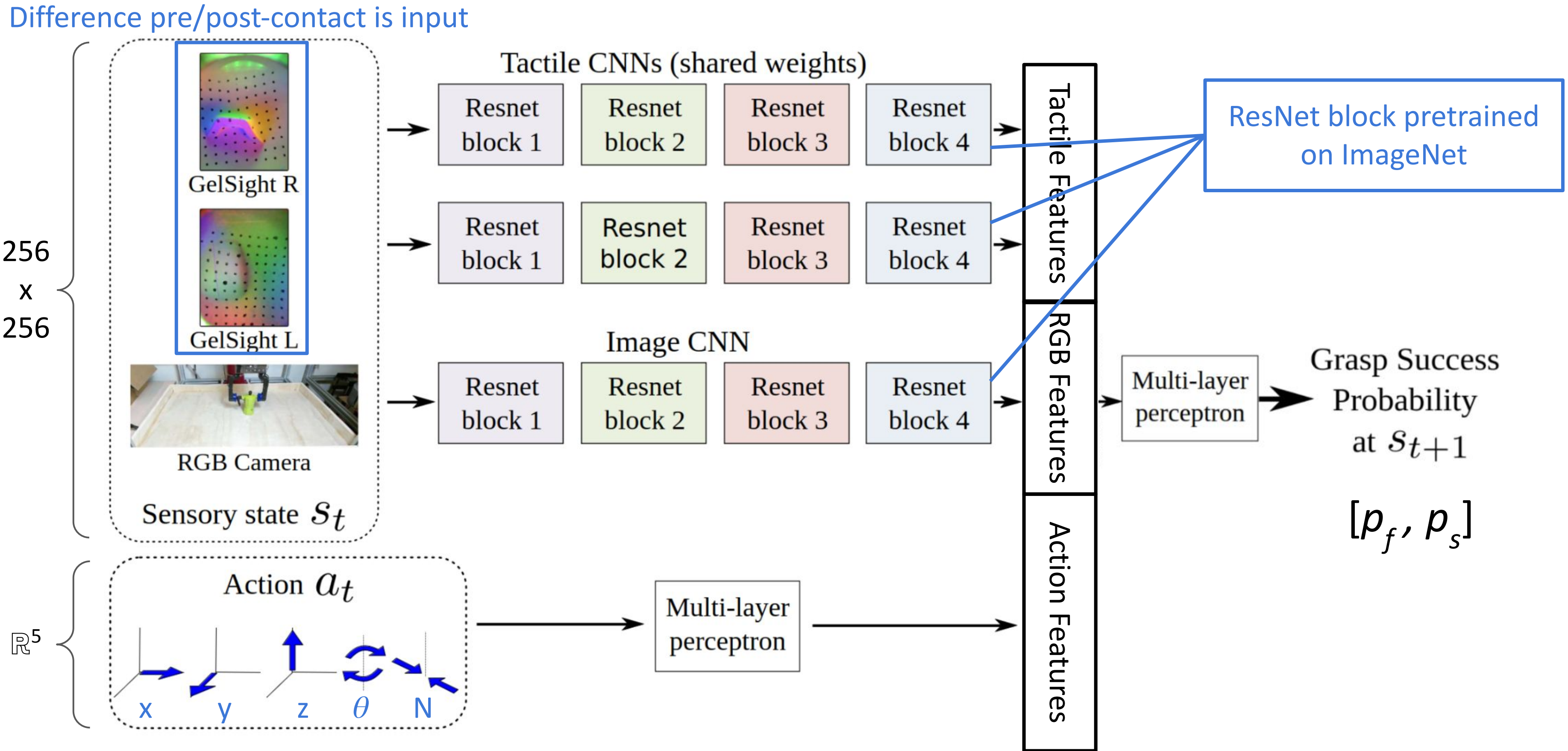
Network Design: Action-conditioned model ***f*(*s*,*a*)**



Network Design: Action-conditioned model *f(s,a)*



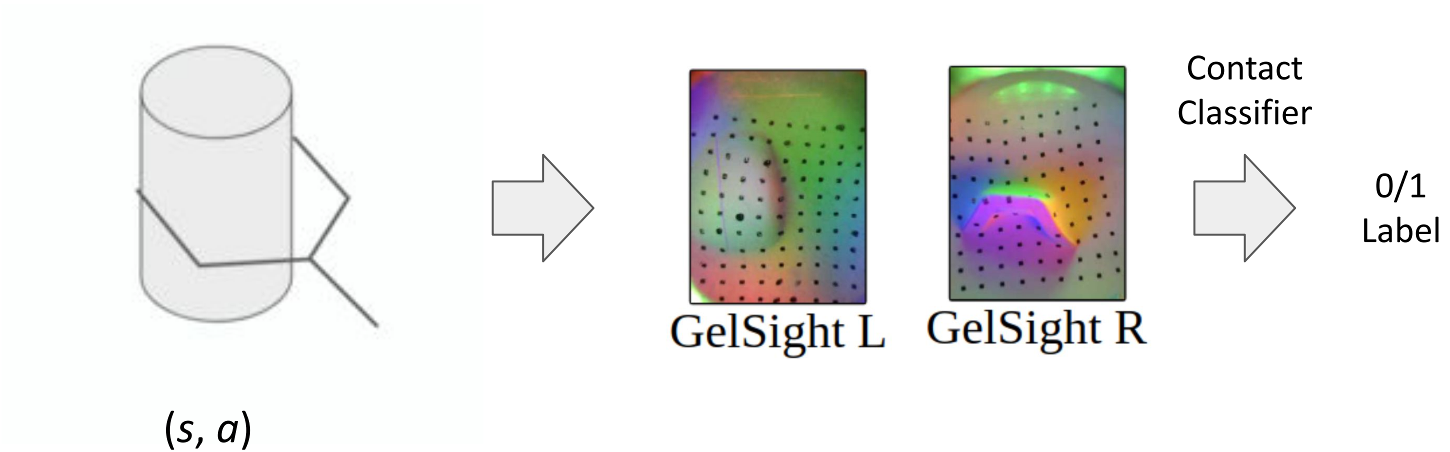
Network Design: Action-conditioned model ***f*(*s*,*a*)**



Data Collection for Self-Supervised Action Outcomes

To collect state-action pairs

- Obtain 3D enclosure of object with depth
- Position grip at center **+ noise** (⟵ action)
- Attempt lift + hold
  - GelSight classifier trained to identify contact at the end of 4 seconds



Comparison on Ablated Data

Q: Does the model base it's predictions on the action i.e. **<u>is it action-conditioned</u>**?

- Essential for test-time iterative grasping

Providing only state s_t diminishes performance

Visuo-tactile information improves over single modality

Model	Accuracy (mean ± std. err.)
Chance	62.80% ± 0.85%
Vision (+ action)	73.03% ± 0.24%
Tactile (+ action)	79.34% ± 0.66%
Tactile + Vision (+ action)	**80.28% ± 0.68%**
Tactile + Vision (no action)	76.43% ± 0.42%

Test-Time Results

Sample for a* with probability > 0.9

$$\boldsymbol{a}_t^* = \arg\ \max_{\boldsymbol{a}} f\left(\boldsymbol{s}_t, \boldsymbol{a}\right)$$

<u>*a* sampled from</u>:

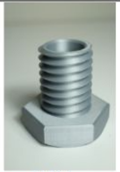
$$\left.\begin{array}{c}[-2, 2]\text{cm } (x,y,z) \text{ translation} \\ [-17, 17]^{O} \text{ rotation} \\ [4, 25]\text{N grasp force}\end{array}\right\} a$$

If $f(s_t, a) > 0.9 \longrightarrow$ lift

Test-Time Results

Sample for a* with probability > 0.9

$$a_t^* = \arg\max_{a} f(s_t, a)$$

"Easy" set	Objects / Methods												Average grasp success
													
		215g	160g	40g	125g	125g	65g	135g	30g	380g	140g	10g	
		% grasp success (# success / # trials)											
	Vision only	76% (38/50)	70% (7/10)	60% (6/10)	50% (5/10)	50% (5/10)	90% (9/10)	40% (4/10)	60% (6/10)	**90% (9/10)**	10% (1/10)	**100% (10/10)**	63.2%
	Tactile + Vision	**95% (95/100)**	**100% (10/10)**	**100% (10/10)**	**100% (10/10)**	90% (9/10)	**100% (10/10)**	**90% (9/10)**	**100% (10/10)**	80% (8/10)	**90% (9/10)**	90% (9/10)	**94.0%**
	Cylinder fitting	90% (18/20)	90% (18/20)	80% (16/20)	55% (11/20)	**100% (20/20)**	**100% (20/20)**	**90% (18/20)**	75% (15/20)	35% (7/20)	20% (4/20)	**100% (20/20)**	75.9%

Retraining with data collected by learned model

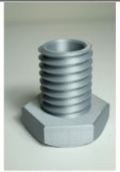
"Hard" set	Objects / Methods												Average grasp success
													
		230g	120g	195g	50g	70g	85g	38g	165g	65g	340g	110g	
		% grasp success (# success / # trials)											
	Vision only	60% (6/10)	80% (8/10)	30% (3/10)	30% (3/10)	80% (8/10)	40% (4/10)	60% (6/10)	50% (5/10)	50% (5/10)	50% (5/10)	20% (2/10)	50%
	Tactile + Vision	80 % (8/10)	**100% (10/10)**	**50% (5/10)**	80% (8/10)	**90% (9/10)**	**70% (7/10)**	**100% (10/10)**	40% (4/10)	**60% (6/10)**	**80% (8/10)**	60% (6/10)	**73.6%**
	Cylinder fitting	**95% (19/20)**	**100% (20/20)**	35% (7/20)	**100% (20/20)**	**90% (18/20)**	15% (3/20)	90% (18/20)	**85% (17/20)**	15% (3/20)	15% (3/20)	**95% (19/20)**	66.8%

Test-Time Results

Sample for a* with probability > 0.9

$$a_t^* = \arg\ \max_{a} f\left(s_t, a\right)$$

“Easy” set

Objects / Methods	 215g	 160g	 40g	 125g	 125g	 65g	 135g	 30g	 380g	 140g	 10g	Average grasp success
	% grasp success (# success / # trials)											
Vision only	76% (38/50)	70% (7/10)	60% (6/10)	50% (5/10)	50% (5/10)	90% (9/10)	40% (4/10)	60% (6/10)	**90% (9/10)**	10% (1/10)	**100% (10/10)**	63.2%
Tactile + Vision	**95% (95/100)**	**100% (10/10)**	**100% (10/10)**	**100% (10/10)**	90% (9/10)	**100% (10/10)**	**90% (9/10)**	**100% (10/10)**	80% (8/10)	**90% (9/10)**	90% (9/10)	**94.0%**
Cylinder fitting	90% (18/20)	90% (18/20)	80% (16/20)	55% (11/20)	**100% (20/20)**	**100% (20/20)**	**90% (18/20)**	75% (15/20)	35% (7/20)	20% (4/20)	**100% (20/20)**	75.9%

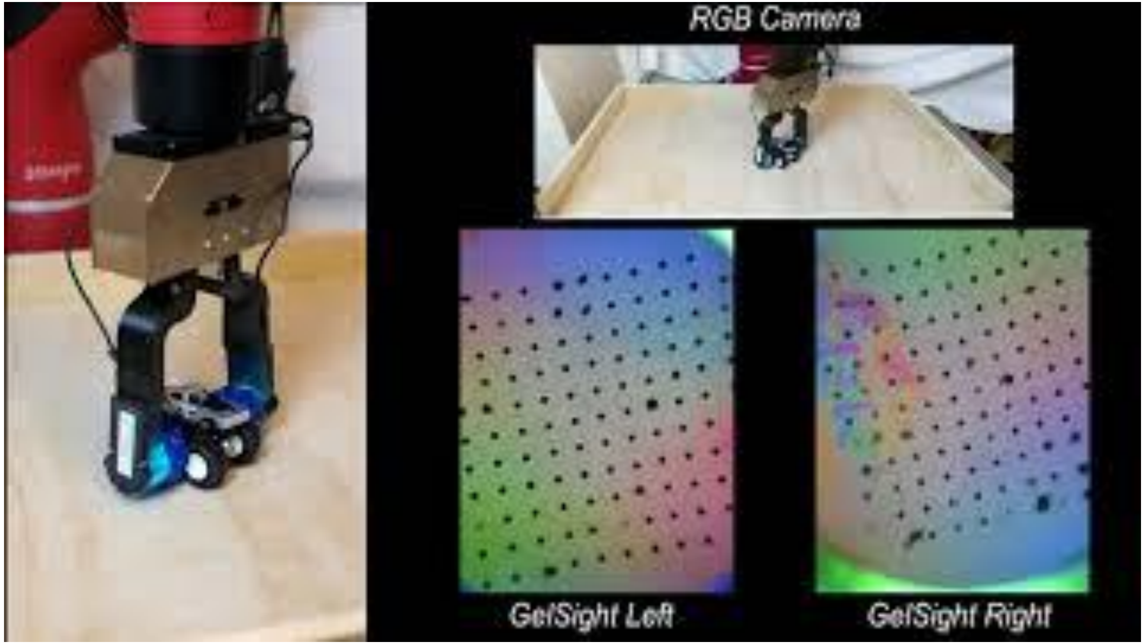
Retraining with data collected by learned model

“Hard” set

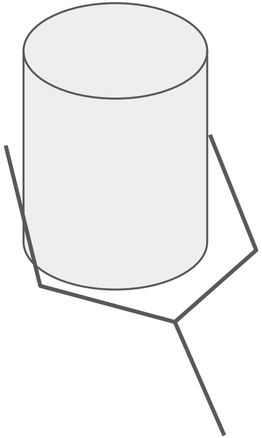
Objects / Methods	 230g	 120g	 195g	 50g	 70g	 85g	 38g	 165g	 65g	 340g	 110g	Average grasp success
	% grasp success (# success / # trials)											
Vision only	60% (6/10)	80% (8/10)	30% (3/10)	30% (3/10)	80% (8/10)	40% (4/10)	60% (6/10)	50% (5/10)	50% (5/10)	50% (5/10)	20% (2/10)	50%
Tactile + Vision	80 % (8/10)	**100% (10/10)**	**50% (5/10)**	80% (8/10)	**90% (9/10)**	**70% (7/10)**	**100% (10/10)**	40% (4/10)	**60% (6/10)**	**80% (8/10)**	60% (6/10)	**73.6%**
Cylinder fitting	**95% (19/20)**	**100% (20/20)**	35% (7/20)	**100% (20/20)**	**90% (18/20)**	15% (3/20)	90% (18/20)	**85% (17/20)**	15% (3/20)	15% (3/20)	**95% (19/20)**	66.8%

Most improvement seen on deformable objects or visually difficult objects

Video Demonstration



Examining the Learned Model: Grasping Force

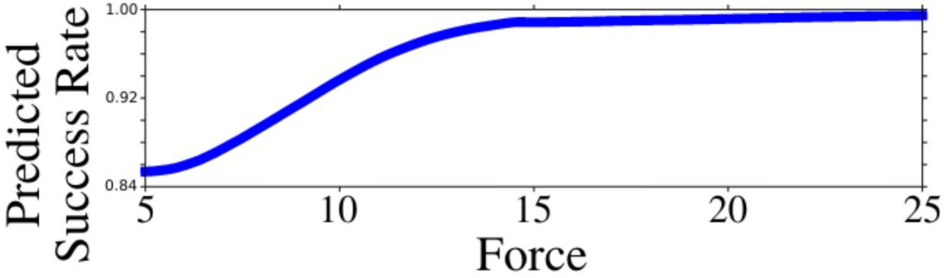


Fix s_t

Enumerate grip force among actions

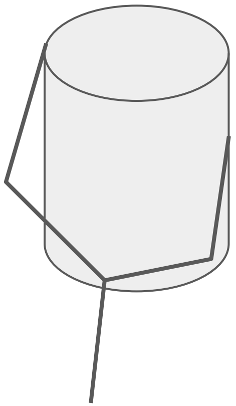
a_1	a_2	a_3	a_4	5
a_1	a_2	a_3	a_4	6
a_1	a_2	a_3	a_4	7
a_1	a_2	a_3	a_4	8

⋮



[Stable Grip]

Examining the Learned Model: Grasping Force

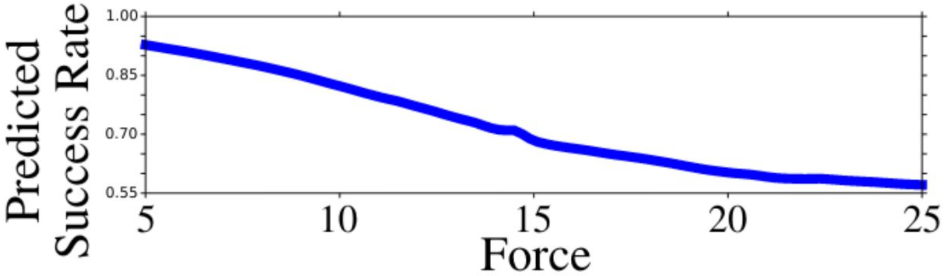


Fix s_t

Enumerate grip force among actions

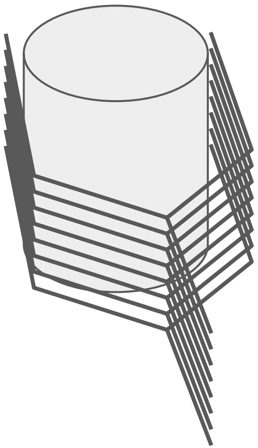
a_1	a_2	a_3	a_4	5
a_1	a_2	a_3	a_4	6
a_1	a_2	a_3	a_4	7
a_1	a_2	a_3	a_4	8

⋮

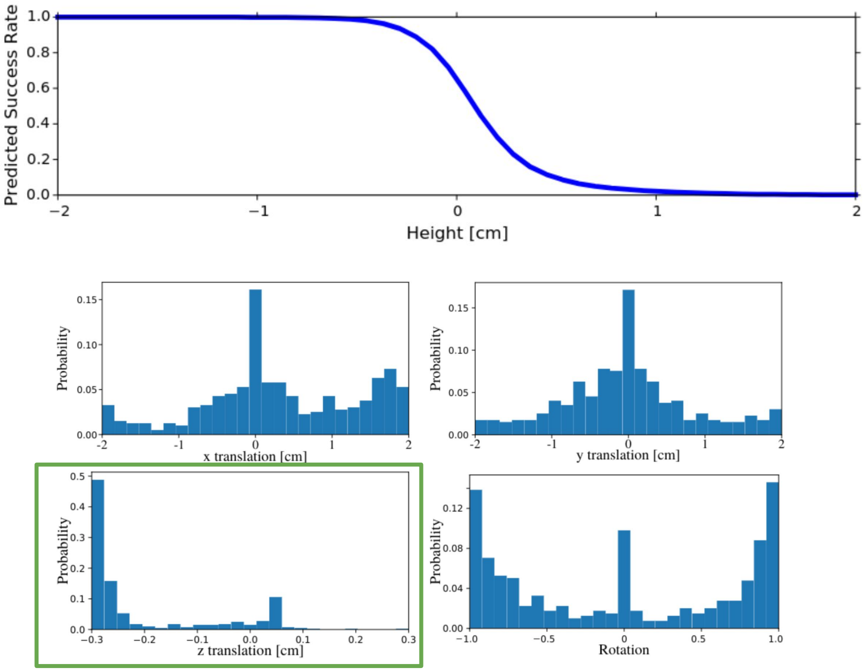


[Unstable Grip]

Examining the Learned Model: Grip Height



Position grip



Across trials, decreasing height of gripper leads to higher success rates

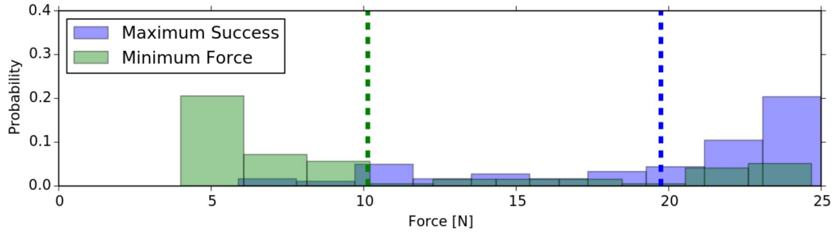
Examining the Learned Model: Minimal Force Grasping

Change objective to minimize force (a_5)

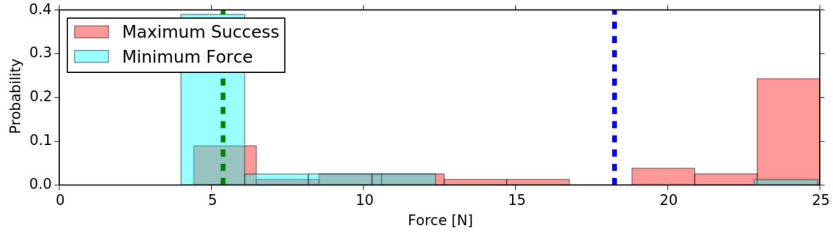
$$\boldsymbol{a}_t^* = \arg\,\max_{\boldsymbol{a}} f\left(\boldsymbol{s}_t, \boldsymbol{a}\right)$$

$$a_t^* = \mathrm{argmin}_a a_5$$
$$\text{s.t.} \quad f(s_t, a) > 0.9$$

Performance with minimum force optimization matches maximum success



(a) Tactile+Vision



(b) Vision only

Future Directions

Limitations

- No information gathering - single-step predictions
- Coarse actions
- No realistic, varied environments

Summary

- Visual + Tactile
- Regrasping
- Data-driven

Discussion

@111_f1 Tactile information can provide benefits such as reducing the amount of force applied to the object while maintaining a secure grip. A multimodal approach is a promising step **towards more human-like robotic grasping**, which could improve overall performance and utility in the real-world.

Q: **Do we think the human-inspired model design trend is useful?** In what ways? What could it be applied to?

- Alternative question is “what is the alternative?”
- Seen before (emergence of objectness, body schema, learning affordances)

Discussion

@111_f5 Tactile feedback helps humans grasp/lift (and adjust) successfully, which this model incorporates.

Still, humans can know how to grasp new objects by seeing them, and **incorporating general knowledge** about the world e.g. a toy car will be light and have spin-able wheels.

Q: In many cases, we need no tactile data and next-to-no visual examination to successfully lift - why?

- We may have some sort of foundational tactile prior, which would be very difficult to obtain in systems in the same way we have seen foundational models work for text or image data. There is far fewer usable manipulation/tactile data for this case.
- We also have good mid-level representation which allows us to generalize

Piazza Discussion

Benefits of GelSight Sensor and tactile feedback:

- Takes advantage of feedback that human gets in grasping and balancing object to grasp better and even delicately (when optimized for minimal force)
- Con of sensor is it wears quickly - even if model can become invariant to this, it still takes manual replacement
- Interesting to see how this tactile data could be used to react/adjust post-grasp

Piazza Discussion

How to adjust/react to disturbances and shifts in grasp:

- Temporally dense proprioceptive and tactile feedback (as opposed to open-loop control post-grasp) could be used to adjust in real-time as seen in legged robotic literature
- No robustness to potential disturbances which would knock object or robot and compromise grip
- Could extend "regrasping" idea to post-grip process
- Raw reaction time of system could remain a difficult problem

Piazza Discussion

Self supervision

- Extremely convenient/important that we can obtain labeled data for manipulation task automatically - tedious and time consuming otherwise
- Useful for scaling

Generalization to new objects

- Not ideal to have to retrain to expand functionality to “hard” test set
- Accuracy/performance seems to rely largely on tactile sensor (when comparing vision, tactile, and tactile + vision), which seems more generalizable than vision