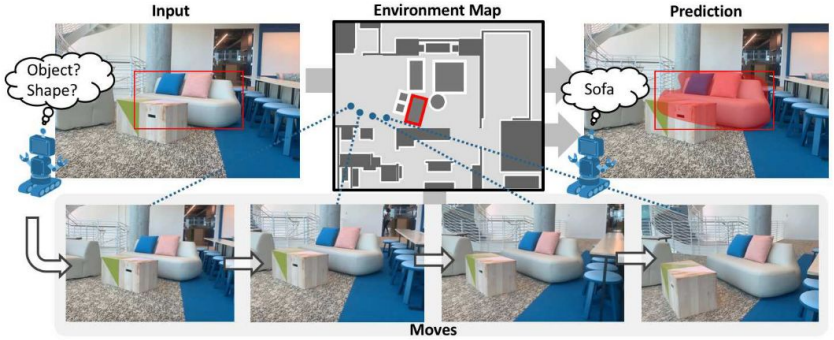
Embodied Amodal Recognition: Learning to Move to Perceive Objects

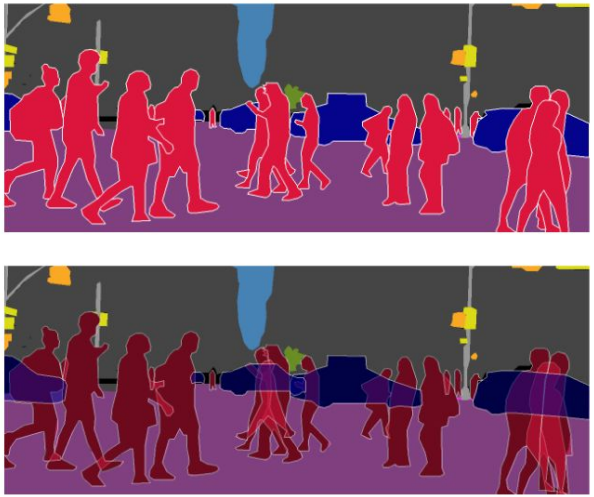
Presenter: Annabel To

1 September 2022

Problem

- At the moment, robots lack a model for embodied amodal recognition, or the ability to reposition and classify an entire object that was initially occluded.
    - Current amodal recognition: struggles on still images, hallucinating (3D framing → 2D shape)
- Big idea: how does movement and accrued perspectives in a 3D environment affect object localization, segmentation, and recognition?





Applications & Significance

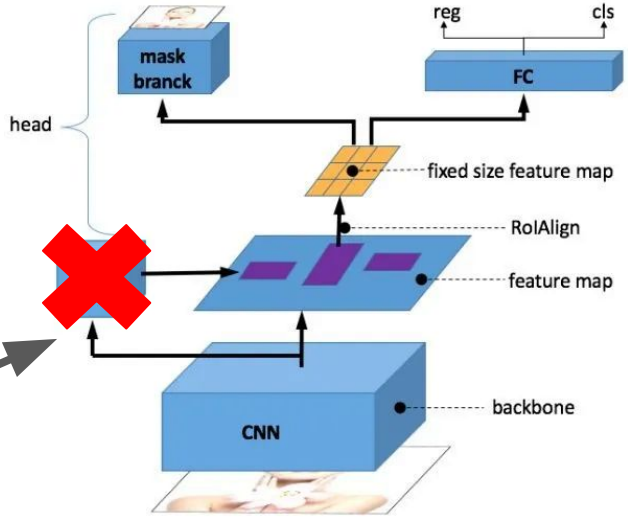
- Why should we care?
  - deployment in the real world: improving adaptability for dynamic and unstructured environments
  - challenge of occlusion reasoning (depth, dimension) remains despite great progress in image segmentation

Current Work

- "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." Robert Geirhos et. al.
  - Common method: creates 3D frame and projects frame onto 2D shape
  - Complex image recognition task → overfitting, expensive
- "Semantic Amodal Segmentation." Yan Zhu et. al
  - Amodal annotations → 3D understanding of world, object permanence
    - Annotations resolve uncertainties on occluded portion (or whole)
  - Implicit deep learning of 3D shape causes overfitting to dataset

Implementation

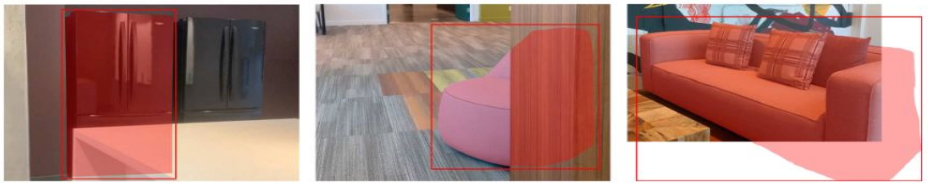
Implementation Overview



- Restrict focus to single-target object's visible region
  - spawn in and capture initial bbox
- Recognition network: Mask R-CNN w/ ResNet backbone
  - pre-trained on ImageNet
  - replace region proposal (RPN) w/ bbox
- Use of bbox in tandem w/ segmentation mask
  - "Detector Algorithms of Bounding Box and Segmentation Mask of a Mask R-CNN Model" by Haruhiro Fujita et. al.
    - allows learning along detail gradient
- Goal: improving initial amodal recognition (ie. classification, bbox, mask at t=0)
  - attempt amodal recognition right out the gate

Dataset

- Simulated indoor environment; 550 houses total
  - single target, rigid objects "close" to spawn
    - if not rigid, then orientation of occluded portion could be variable
  - recognition module: ground truth annotations provided
  - action policy network: no ground truth provided (~shortest path)
    - shortest path viewpoint dataset further used in training during Stage 1



"easy", "hard", and out-of-frame IRL annotations

Definitions

- Goals: object recognition (classify), 2D amodal localization (bbox), 2D amodal segmentation (mask)
- I_0: initial (spawning) location observation
- b_0: bbox of target's visual portion, viewed from spawn
- π: action policy dictating movement a_t → observes new image I_t w/ view angle v_t
- $y_t = \{c_t, b_t, m_t\}$: category prediction, amodal bbox, amodal mask for first frame
- $y^* = \{c^*, b^*, m^*\}$: true object category, amodal bbox, amodal segmentation mask
    - ideally, achieve this at time step 0

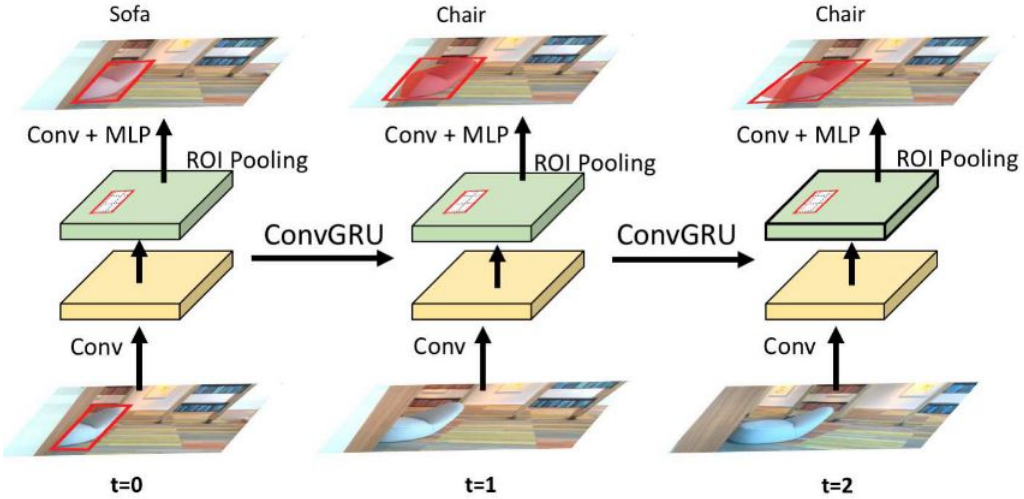
$$L = L_{cls} + L_{box} + L_{mask}$$

$$L^p = \frac{1}{T}\sum_{t=1}^{T}\left[L_c^p(c_t, c^*) + L_b^p(\boldsymbol{b}_t, \boldsymbol{b}^*) + L_m^p(\boldsymbol{m}_t, \boldsymbol{m}^*)\right]$$

loss for amodal recognition, bot (top: Mask R-CNN)

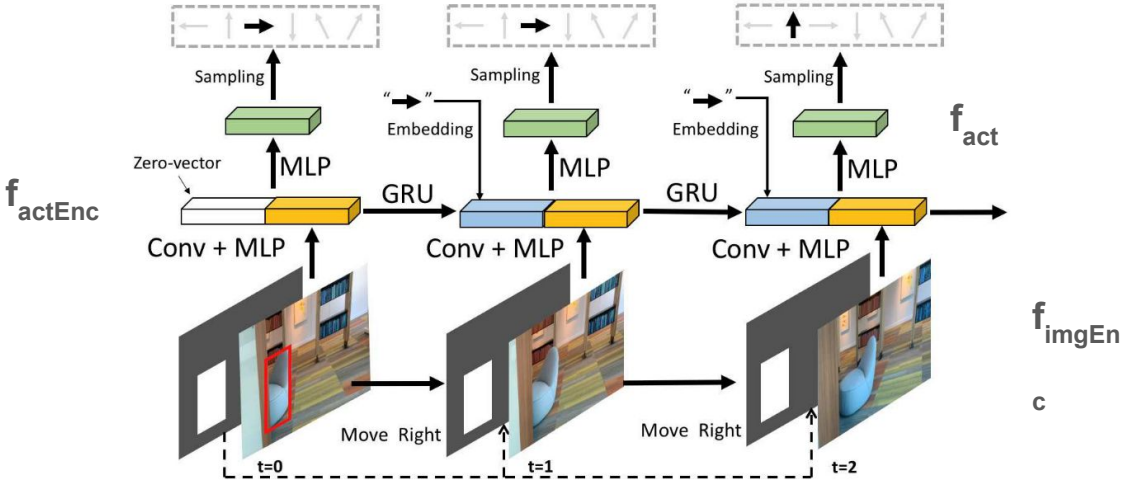
Implementation: Amodal Recognition (Perception)

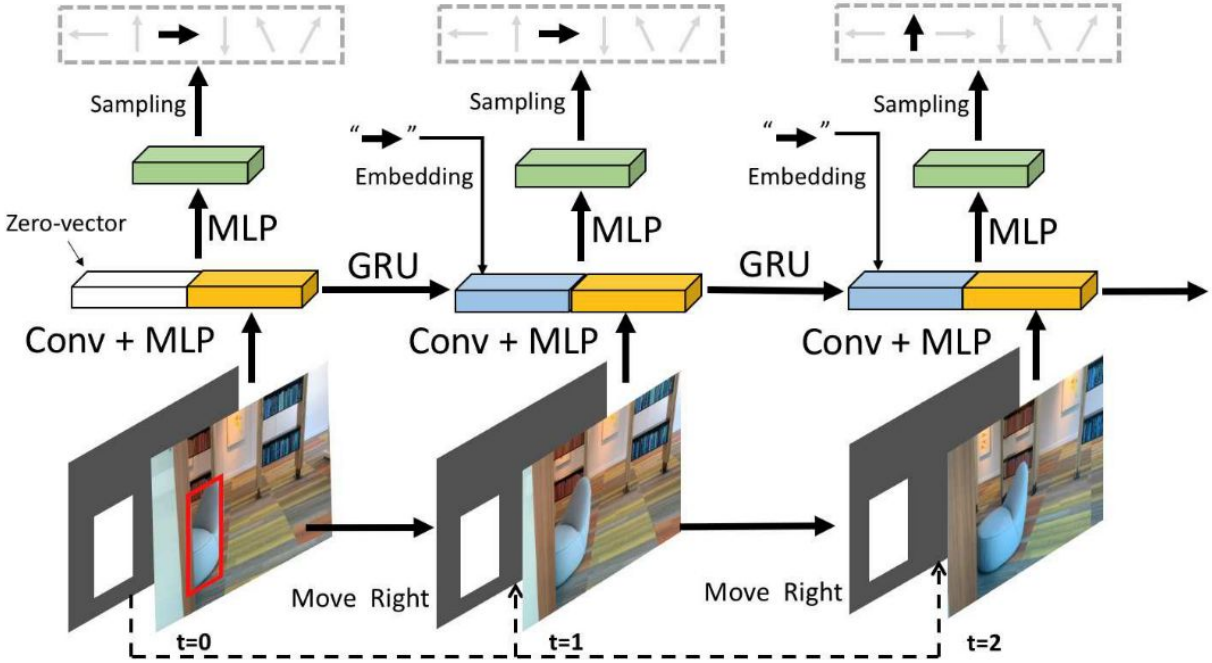
- $f_{base}(I_0)$ encapsulates first frame input
- each frame up to t: aggregated into f_{fuse} (Con-GRU; alternatively pooling or temporal average)
  - Convolutional Gated Recurrent Unit incorporates prior frames
- f_{fuse} sent into f_{head} on RoI; makes prediction for first frame



Implementation: Action Policy Network

- f_{imgEnc}: encoder for image features; takes in I_0, I_t, and mask I^b (representing b_0)
- f_{actEnc}: encodes last action at time t, thus $z_t^{act} = f_{actEnc}(a_{t-1})$
- f_{act}: incorporates prior info (hidden state from last step h_{t-1}) via single-layer GRU network
  - final $z_t = f_{act}([z_{img}, z_{act}], h_{t-1})$ passed to linear layer w/ Softmax over all possible actions





Softmax over all moves

encoding last move a_{t-1}

encoding image input: I_0, I_t, I^b (mask for b_0)

GRU: temporal aggregation, incorporates hidden state h_{t-1}

Architecture for Action Policy Network

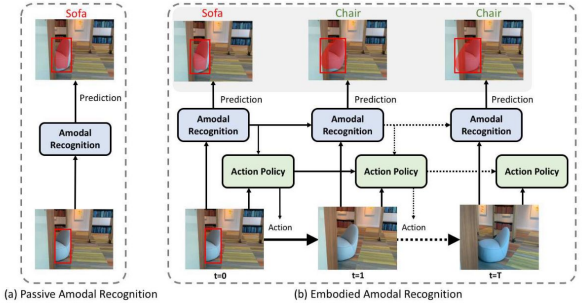
Implementation: Action Rewards

- Reward evaluation:
  - intersection over union for bbox, mask; accuracy for classification
  - weights: $\lambda_c = 0.1$, $\lambda_b = 10$, $\lambda_m = 20$
  - learned via policy gradient method REINFORCE
    - estimates best weights by gradient ascent
    - over trajectories rather than episodes

$$r_t = \lambda_c Acc_t^c + \lambda_b IoU_t^b + \lambda_b IoU_t^m$$
$$R_t = r_t - r_{t-1},$$

reward shaping for action policy

Training

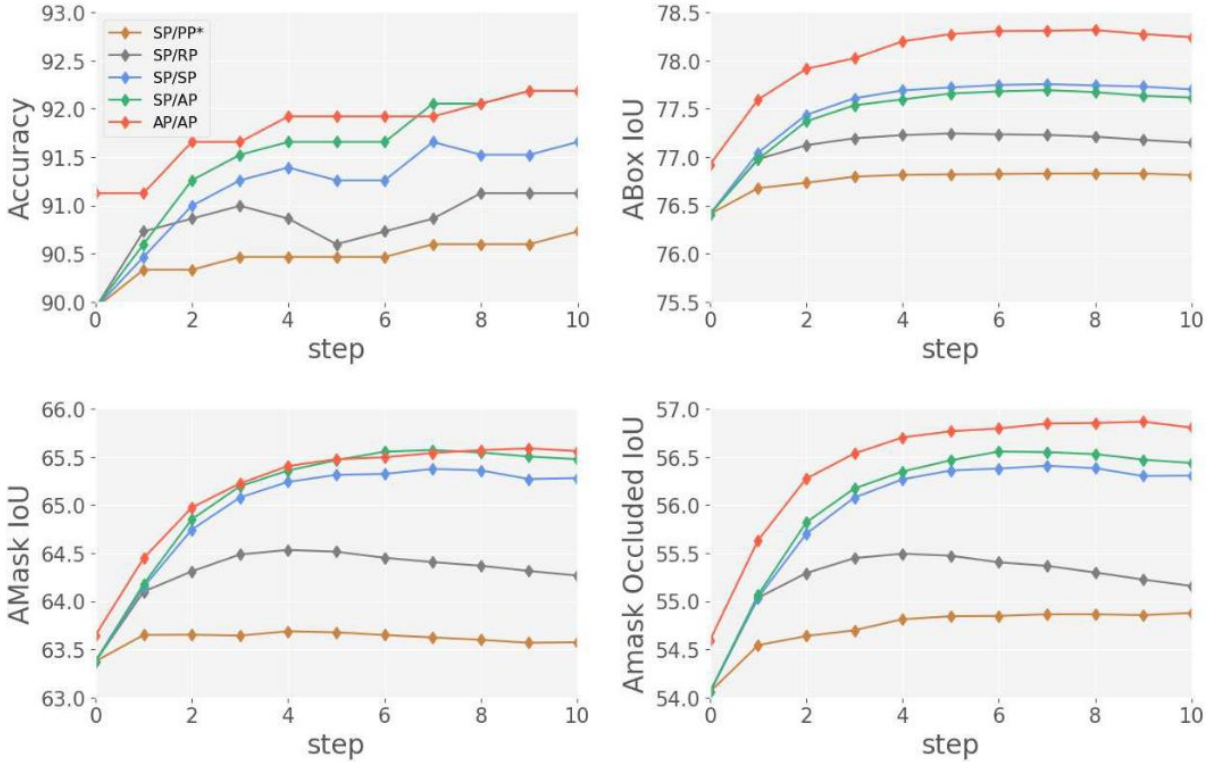


- Goal: separate perception and action module
  - Stage 1: train perception using shortest path images
  - Stage 2: train action policy from scratch using (fixed) input described by perception module
  - Stage 3: continually refine perception module based on movement policy
- Staged training: supports learning along detail gradient
  - "Stage-wise Training: An Improved Feature Learning Strategy for Deep Models" by Elnaz Barshan and Paul Fieguth
- Baselines: combos of passive, shortest path, random path, and active path for training or testing
  - Note: training/testing pairings denote any extra fine-tuning of perception module in Stage 3
  - Final models: ShortestPath/ActivePath (SP/AP), ActivePath/ActivePath (AP/AP)

Results & Analysis

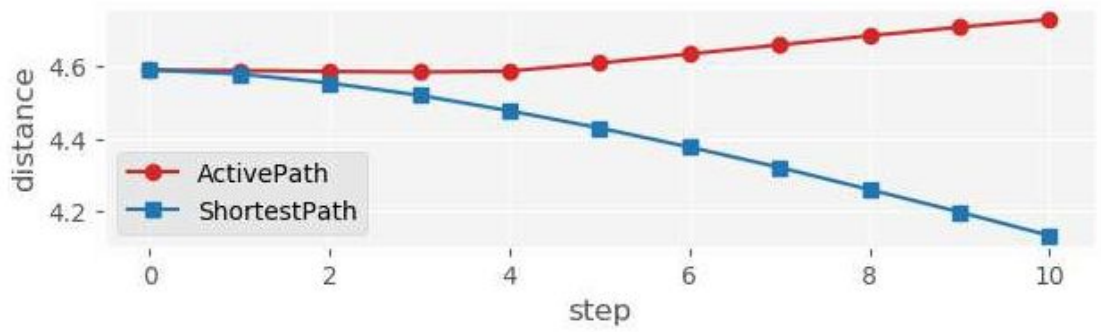
Results

- Key insights
  - Shortest path =/= best path for amodal recognition (passive or active)
  - Embodiment complements amodal recognition; helps with recognizing occlusion
    - overcomes heavy occlusion more readily by a small margin (0.5% higher acc)
  - EAR's staged perception + action policy + perception refinement learns better movement for resolving occlusion
    - non-intuitive movement (eg. backward/orthogonal) learned d.t. post-SP trainings
    - however, as position changes greatly, improvements over step lowers
    - higher performance increase by */AP on "regular," geometric objects



accuracy/step for all methods; movement baseline PP/PP not depicted but performed comparative to SP/PP

Experimental Results



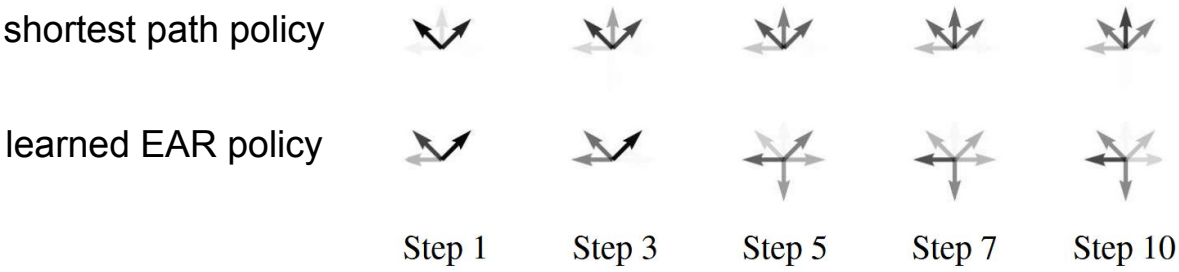
distance to target object; note that shortest path is almost inherently worse in resolving occlusion

relative performance over all methods; baseline (0) is based on worst method in that category

Experimental Results

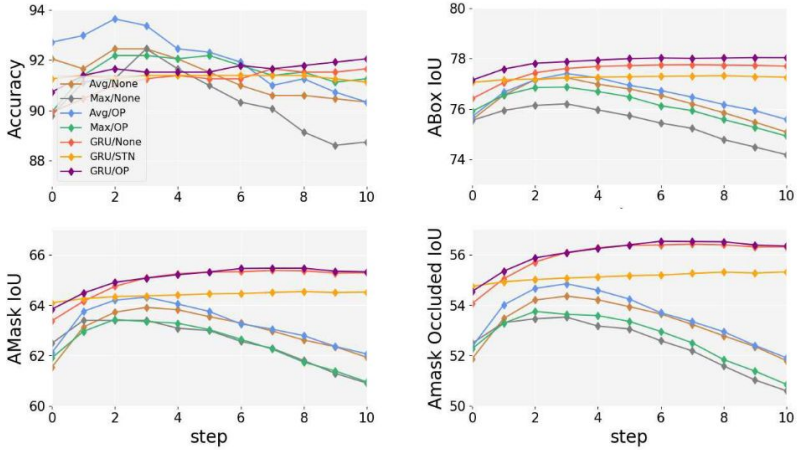
Critiques & Limitations

- Algorithm is restricted to only single target instances and static environments
- Extensive precautions taken to not overfit movement to perception and vice versa
  - staged trainings (3 in total, 2 for perception and 1 for action module) would have to be expanded to generalize as well
- Initial perception training (Stage 1) for final models (SP/AP, AP/AP) is shortest-path
  - further refinement on this baseline is likely to outperform the baseline



Future Expansion

- Expansion for depth ordering and multi-object occlusions/targets
  - Occluding mediums include fog, screens, or lighting differences
- Feature warping and aggregation (Con-GRU replaced/combined with pooling or temporal average): GRU + optical flow had best performance



Further Readings

- On amodal perception/deep learning
  - "Organization in vision: Essays on Gestalt perception." by Gaetano Kanizsa.
  - "Faster r-cnn: Towards real-time object detection with region proposal networks." by Shaoqing Ren et. al.
  - "Semantic Amodal Segmentation." Yan Zhu et. al
- On movement policy in combination with perception module
  - "Amodal completion and size constancy in natural scenes." by Abhishek Kar et. al.
  - "Active object perceiver: Recognition-guided policy learning for object searching on mobile robots." by Xin Ye et. al.
- On staged training/reward shaping
  - "Stage-wise Training: An Improved Feature Learning Strategy for Deep Models" by Elnaz Barshan and Paul Fieguth.
  - "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." Robert Geirhos et. al.

Summary

- Problem: movement to better identify occluded objects
  - opens up new interactions with dynamic, unstructured environments
  - presents embodiment solution to occlusion in deep learning, which most state-of-the-art work attempts to solve with deeper and more complex models
- Key insights
  - successfully learned movement strategy for overcoming occlusion
  - embodiment is crucial to learning optimal movement strategies

Questions for Discussion

1. How would we measure EAR's success regarding multiple (rather than single) targets?
2. The baselines used in this paper combined active, passive, random, and shortest path implementations for both training and testing. What challenges do you foresee in expanding these baselines to account for other agents?
3. How would Stage 1 training (on shortest path images) affect comparative performance to shortest path movement? What if we replaced it with another baseline, ie. radial movement w.r.t. the occluded object?
4. What would be the next steps towards a fully adaptable agent, especially in a dynamic rather than static environment?
5. In your opinion, how could we better ensure that the two modules (perception and action) are separated and unbiased toward each other?
6. Staged training was used because training from scratch failed to build functional rewards in the policy network. How else could we use to solve this issue of network kickstarting?
7. What tradeoffs exist between this model and an unembodied (but perhaps deeper/more complex) amodal recognition algorithm? How could we improve this model to accommodate?