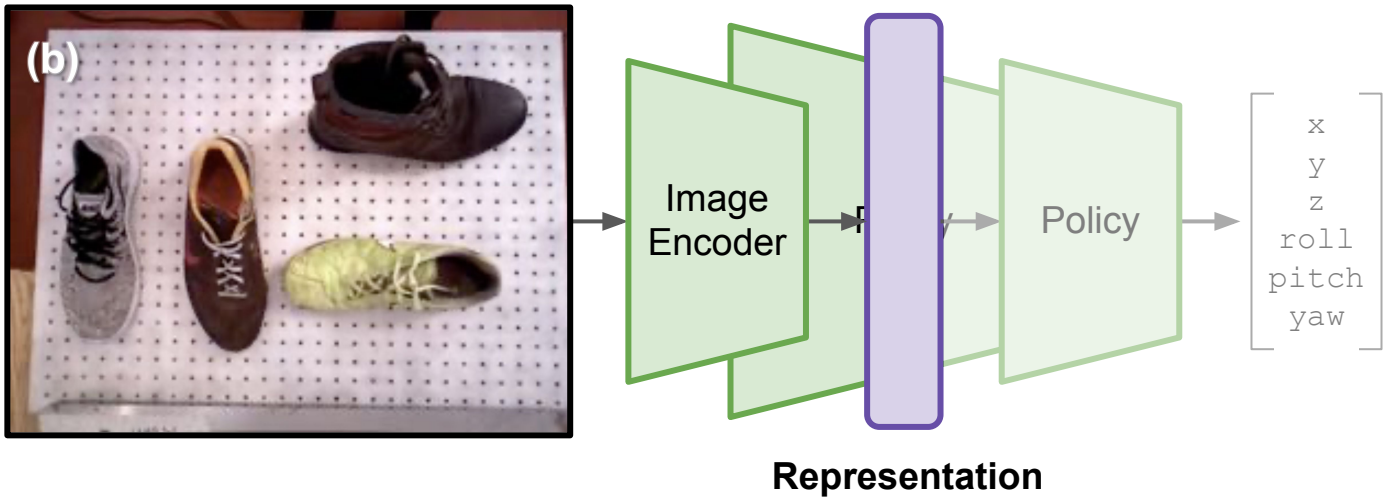
Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation

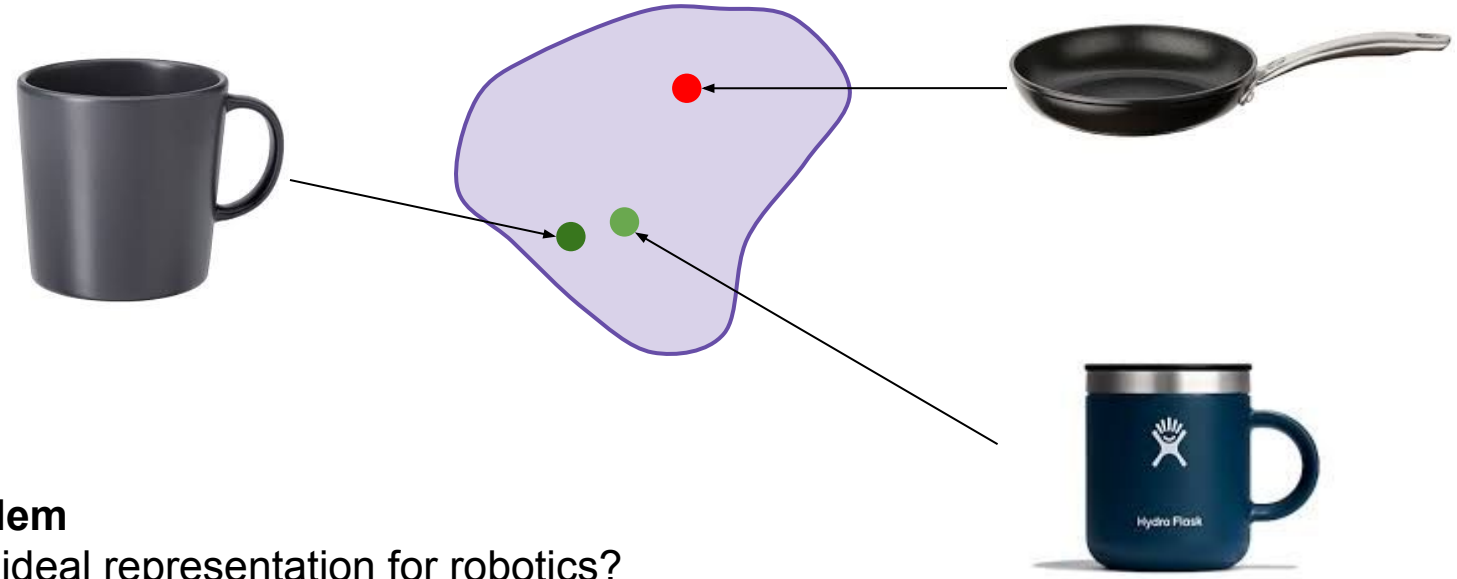
Presenter: Shivin Dass

09-05-2023

How robots perceive their environment?

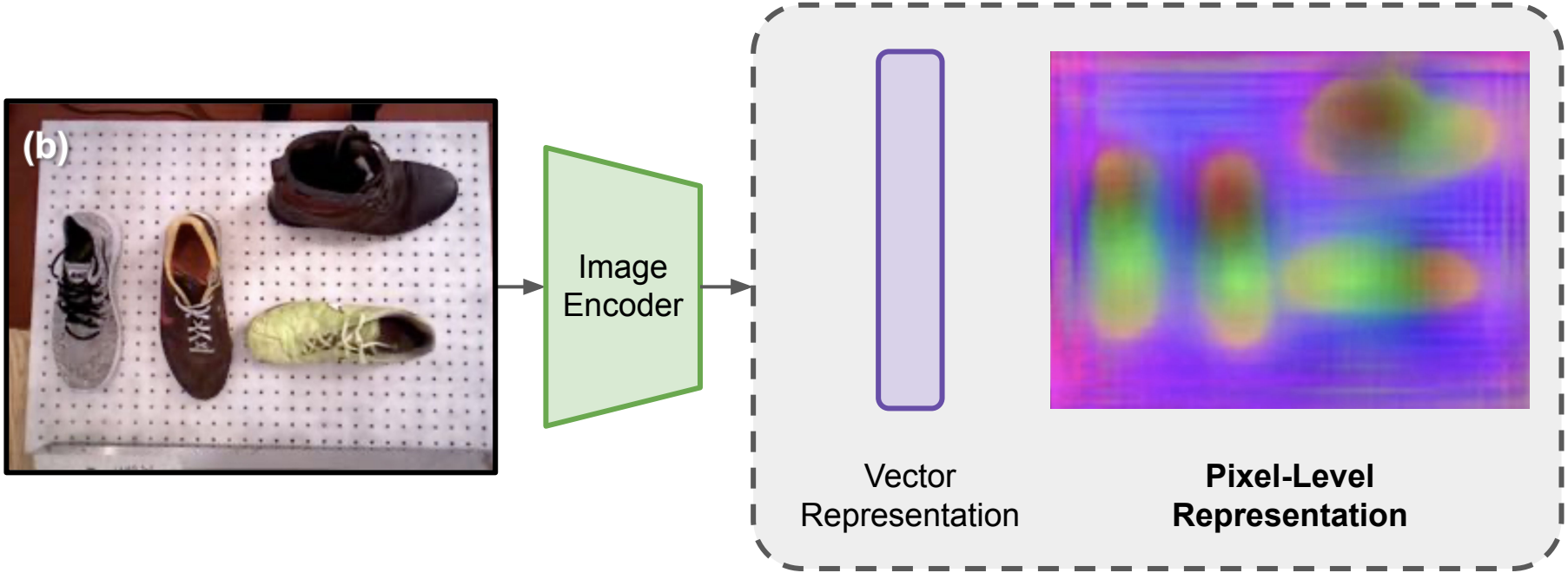


What properties should representations have?

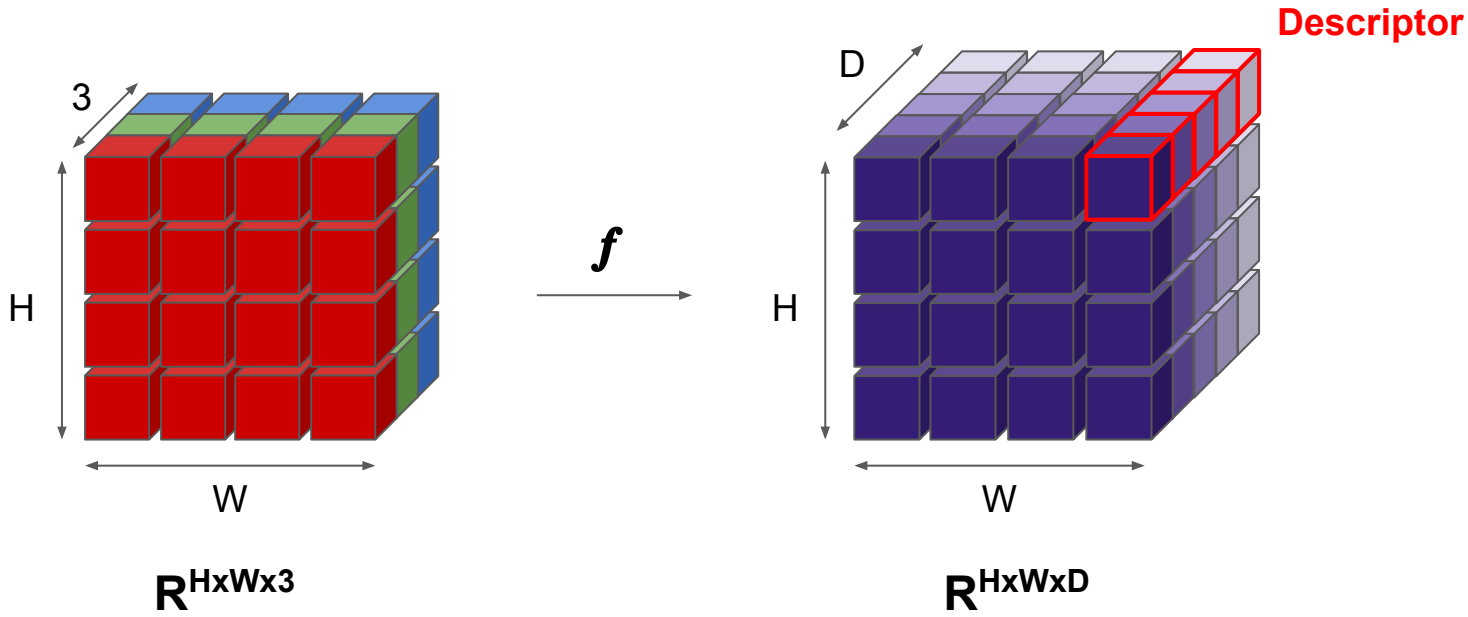


Open Problem
What is the ideal representation for robotics?

Types of representations



Problem Setting

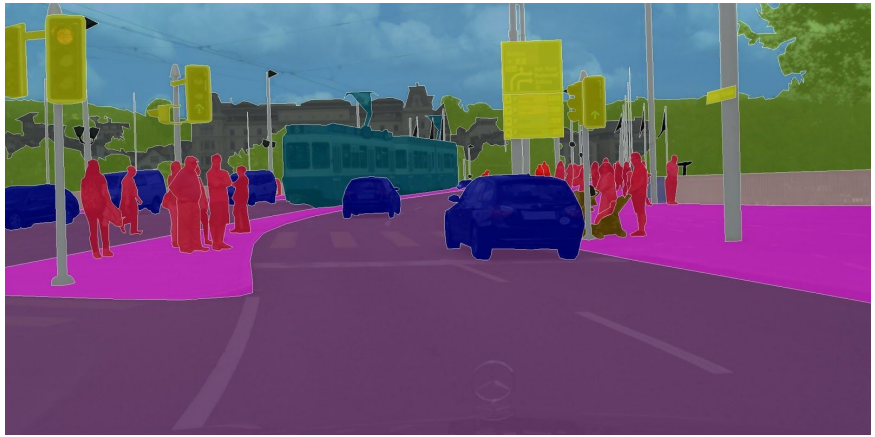


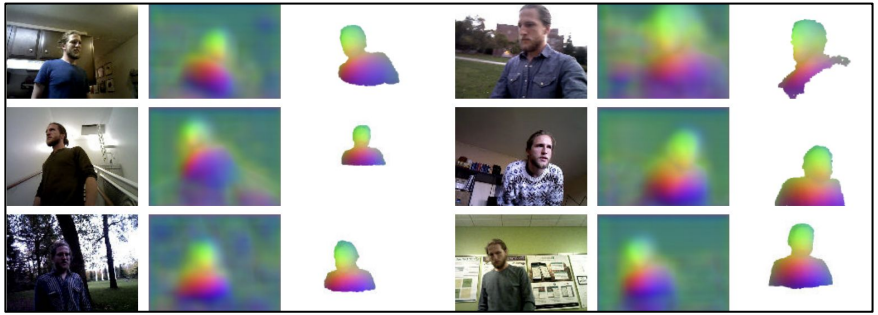
Related Work

- Semantic Segmentation

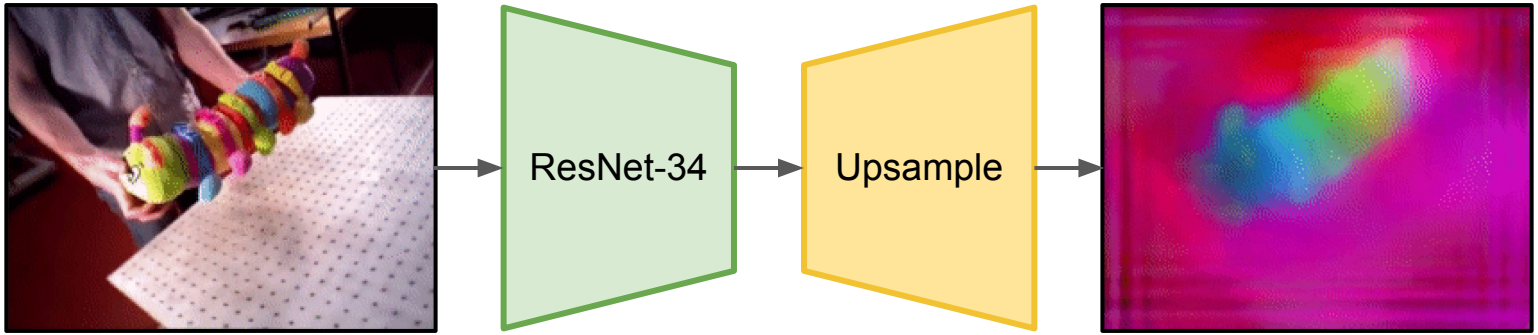
- Learning Dense Correspondences
  - Require labels [1], [2], [3], [4]
  - Not deployed on robots [5], [6]

- Schmidt et al. [6]
  - Uses dynamic reconstruction
  - Single object
  - Not deployed on robots





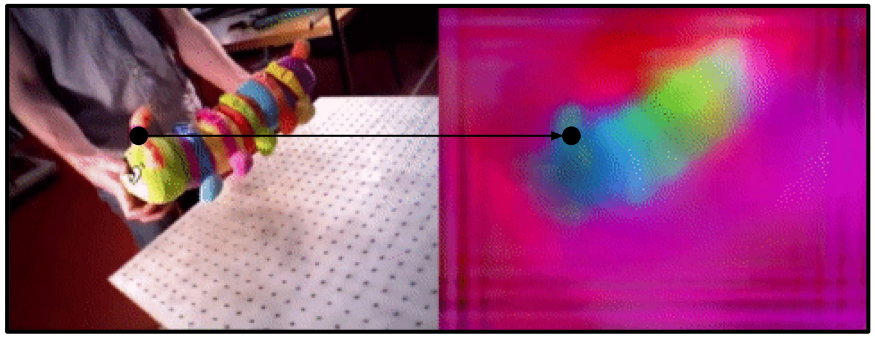
Method: Architecture



But how do we learn the descriptors?

Method: Loss Intuition

Idea: Use a **self-supervised contrastive** objective



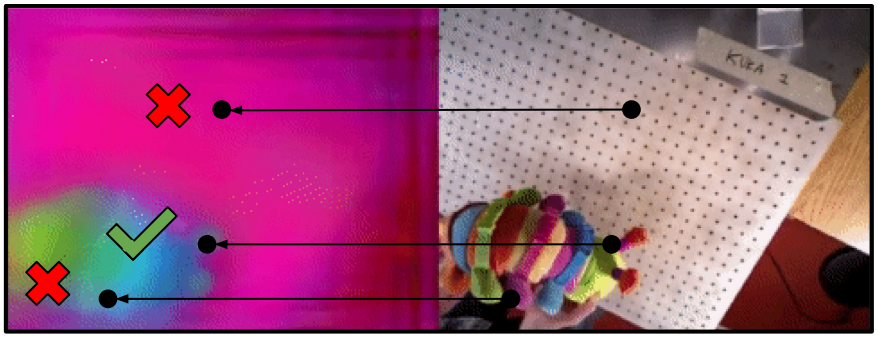


image_1 $f(\text{image}_1)$ $f(\text{image}_2)$ image_2

Matching pairs: Move representations **closer**

Non-Matching pairs: Move representations **apart**

Method: Losses

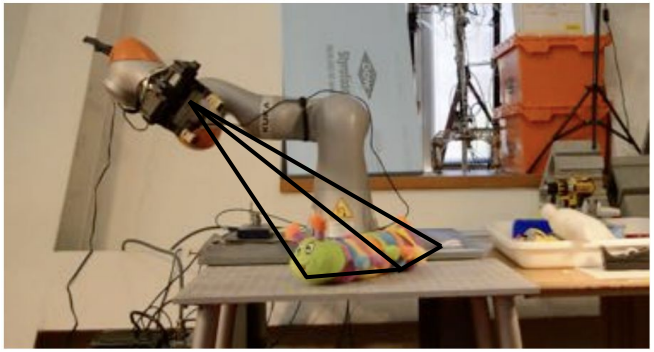
$$D(I_a,u_a,I_b,u_b) \triangleq ||f(I_a)(u_a) - f(I_b)(u_b)||_2$$

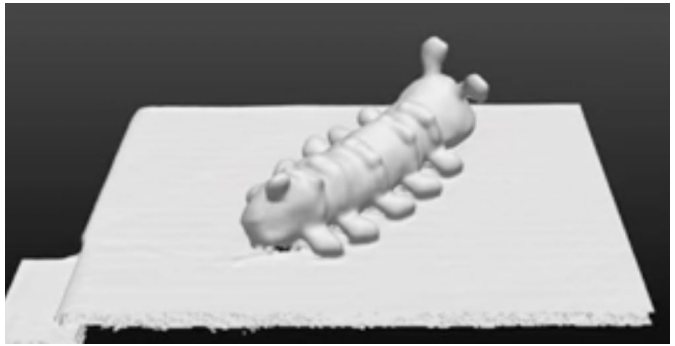
$$\mathcal{L}_{\text{matches}}(I_a,I_b) = \frac{1}{N_{\text{matches}}} \sum_{N_{\text{matches}}} D(I_a,u_a,I_b,u_b)^2$$

$$\mathcal{L}_{\text{non-matches}}(I_a,I_b) = \frac{1}{N_{\text{non-matches}}} \sum_{N_{\text{non-matches}}} max(0, M - D(I_a,u_a,I_b,u_b))^2$$

$$\mathcal{L}(I_a,I_b) = \mathcal{L}_{\text{matches}}(I_a,I_b) + \mathcal{L}_{\text{non-matches}}(I_a,I_b)$$

Method: Dataset Collection



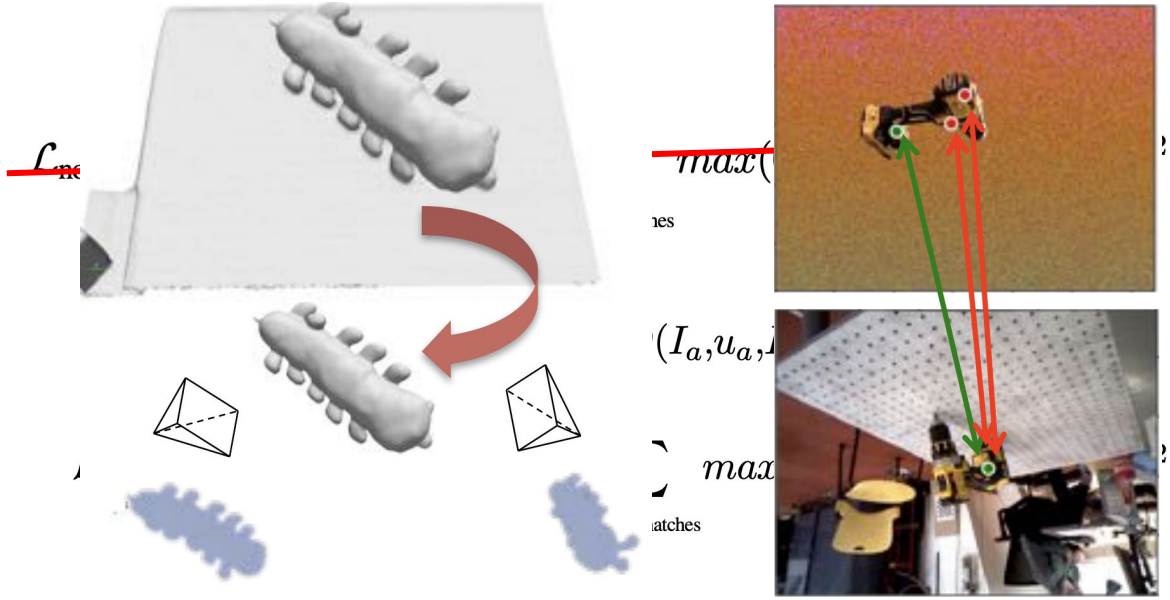


Using **3D reconstruction** and **forward kinematics**, each point can be localized in global frame and correspondences can be created

**use same object pose to collect multiple static scenes

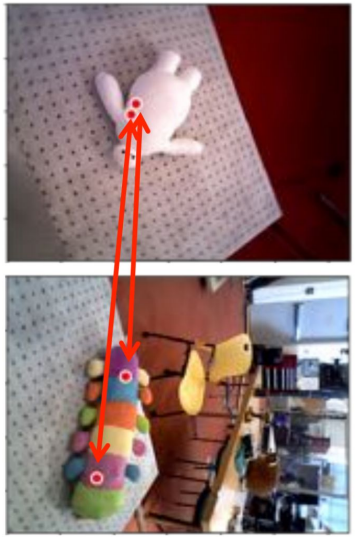
Method: Practical Improvements

- **3D change detection**
- **Background randomization + Data augmentation**
- **Hard negative scaling**

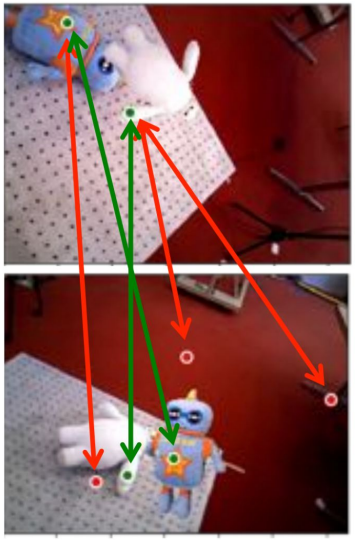


Method: Multi-Object Descriptors

Direct Training on multi-robot scenes



Cross-Object Loss



Synthetic multi-robot scenes

Experimental Setup

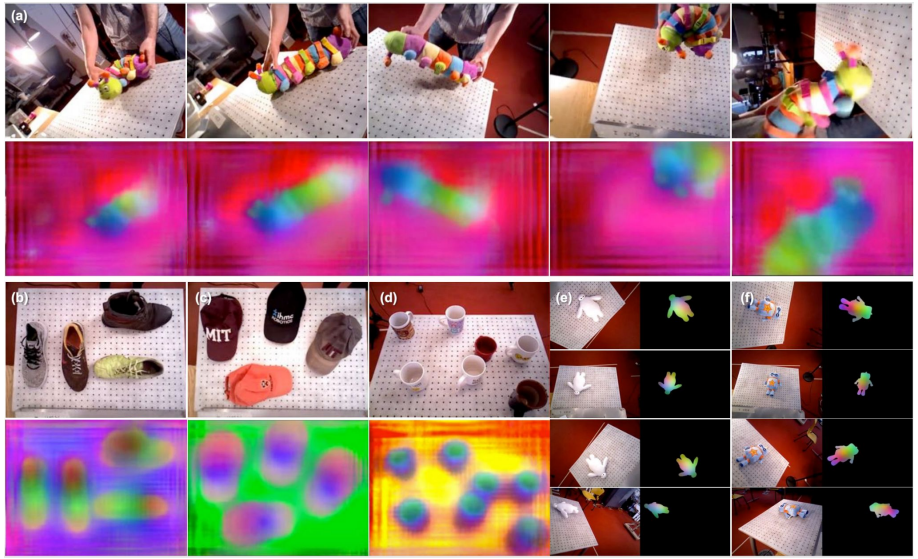
Dataset: Collect own data autonomously*

Robot: Kuka IIWA LRB robot arm

Key Questions

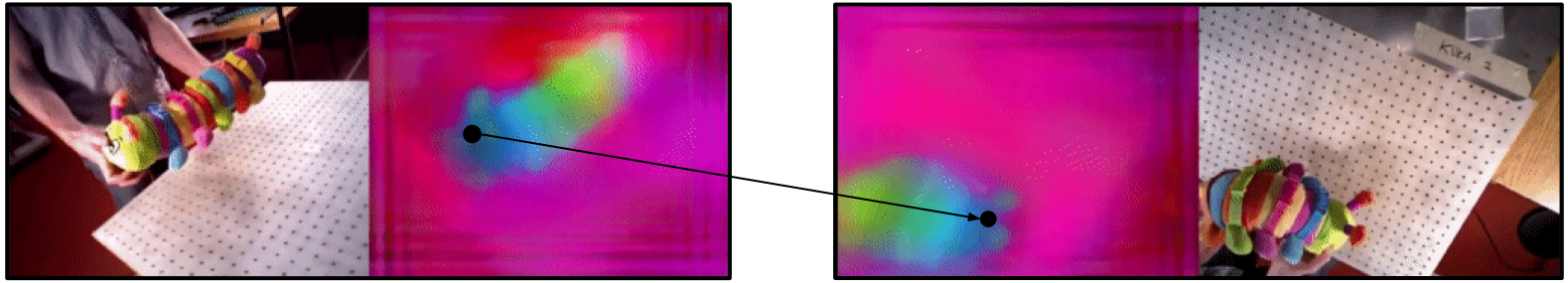
1. Do descriptors learn meaningful representations?
2. What model components contribute the most?
3. Do the descriptors generalize?
4. Can we use descriptors for robot learning?

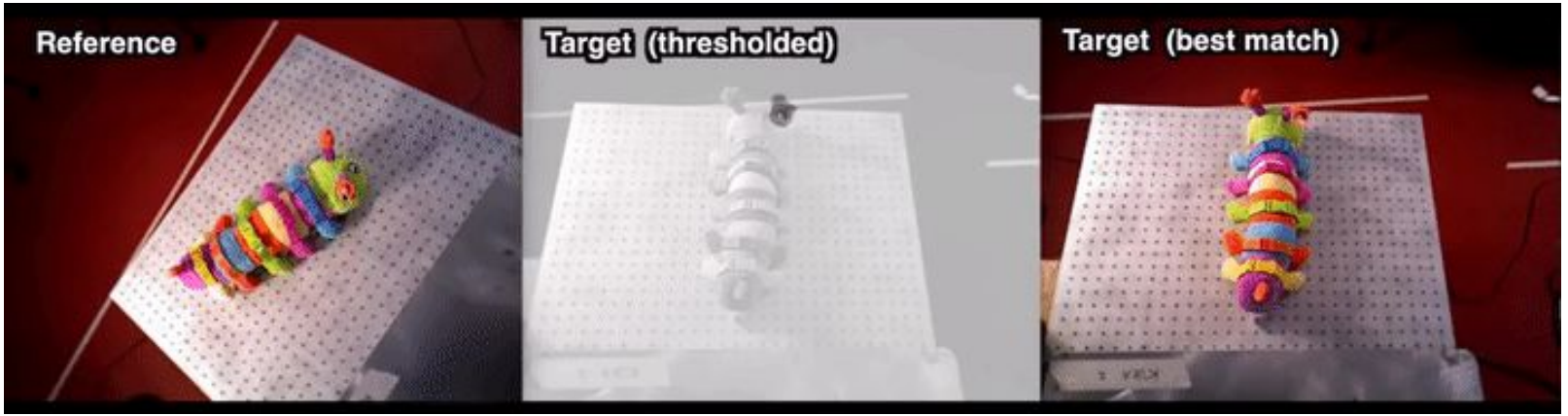
Results



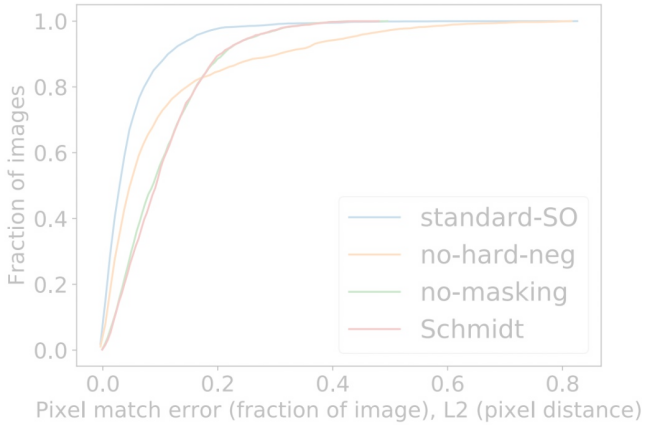
Visualization of learnt descriptors (Fig. 2)

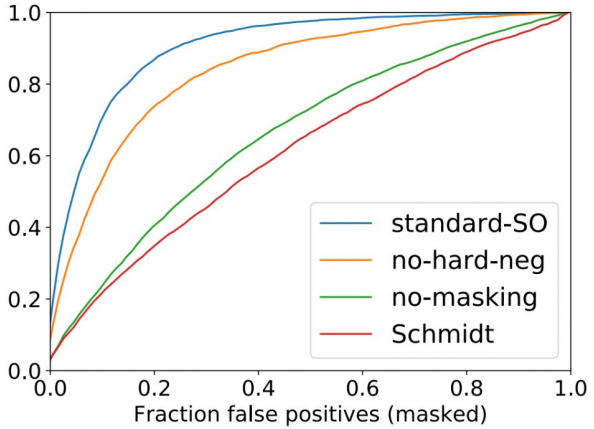
Results: How can we use descriptors?

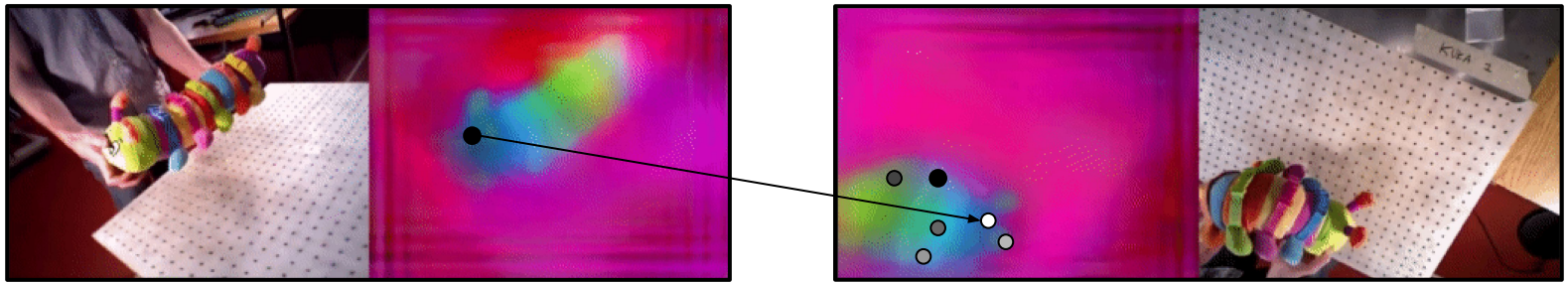




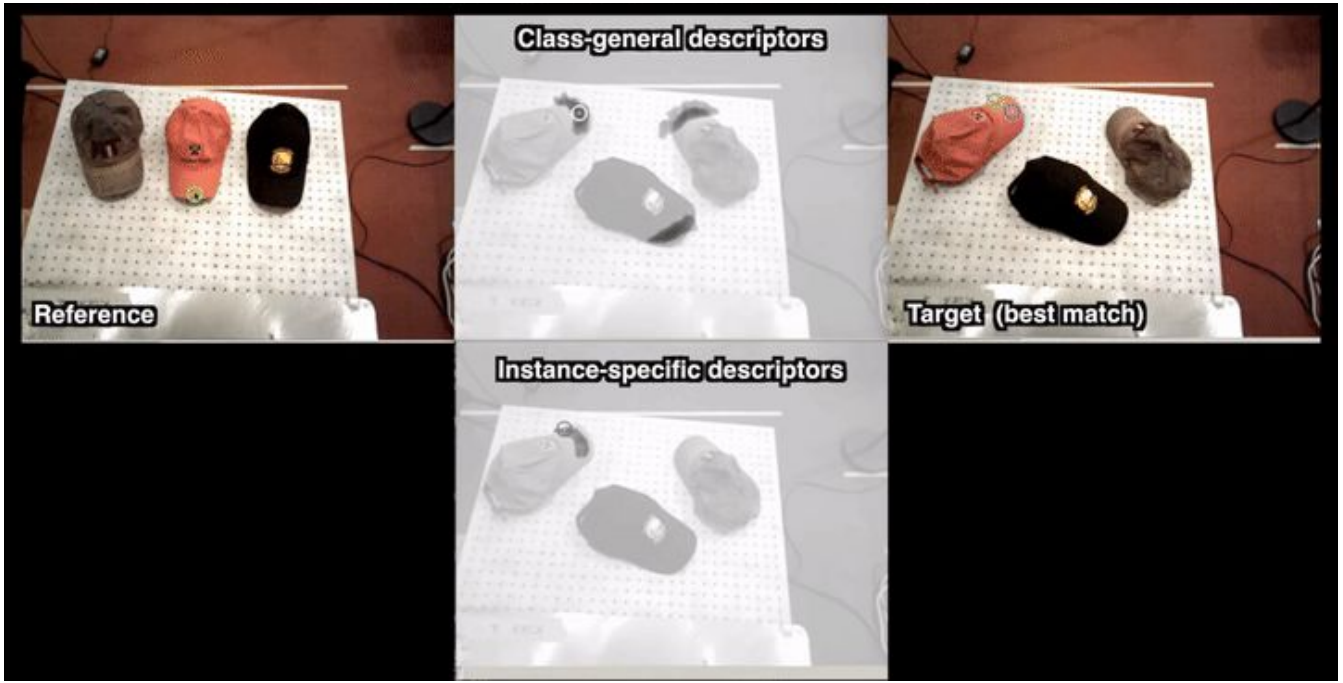
Results: Baseline & Ablations



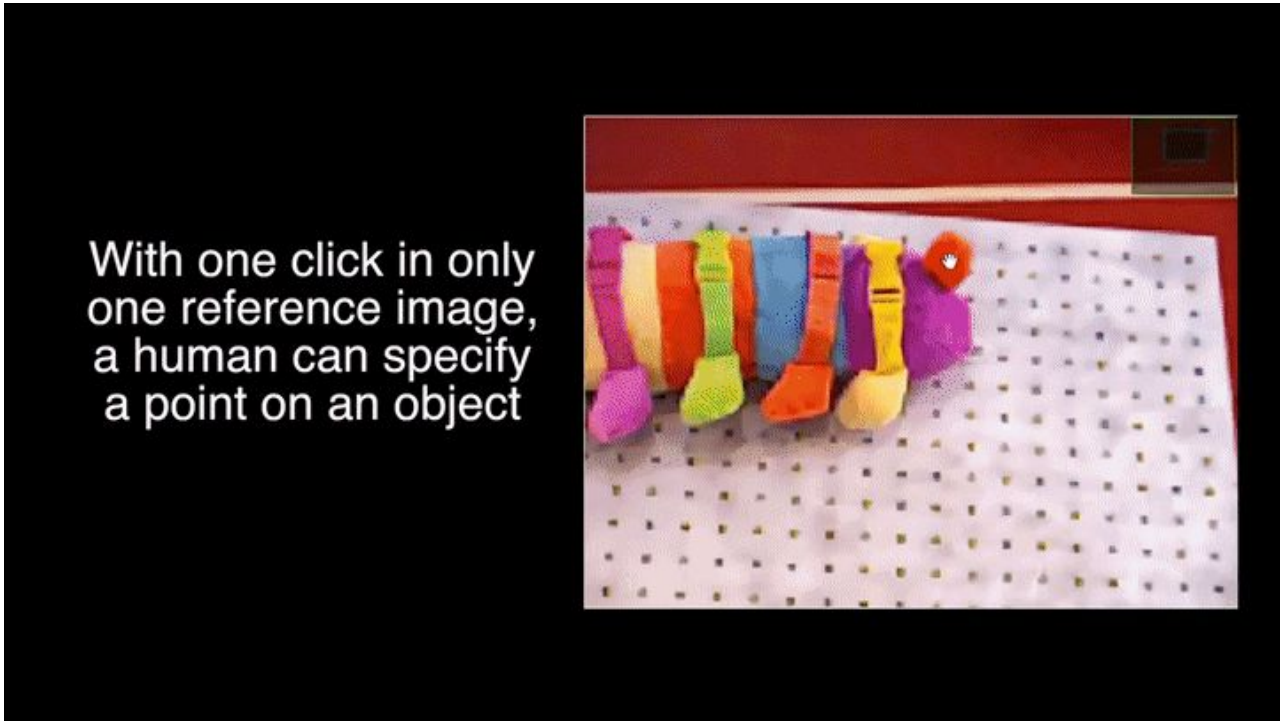




Results: Generalization

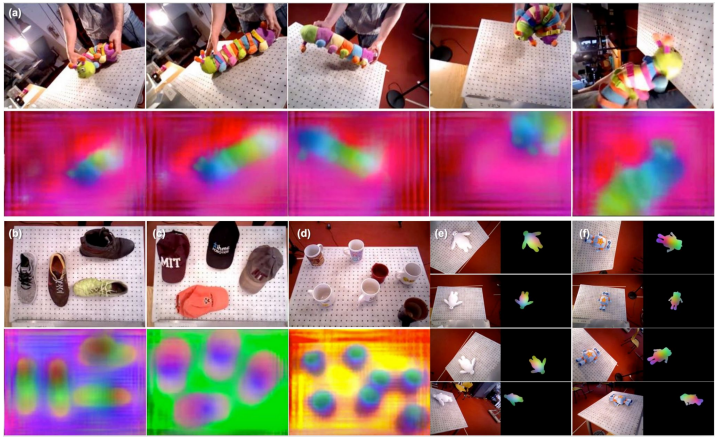


Results: Grasping



Discussion of Results

- **Dense Object Networks** capture object correspondences
- Strong **generalization** performance
  - Position and orientation
  - Class general vs. instance specific
- Can be used for defining target grasps
- Quick to train in ~20 mins!



Limitations

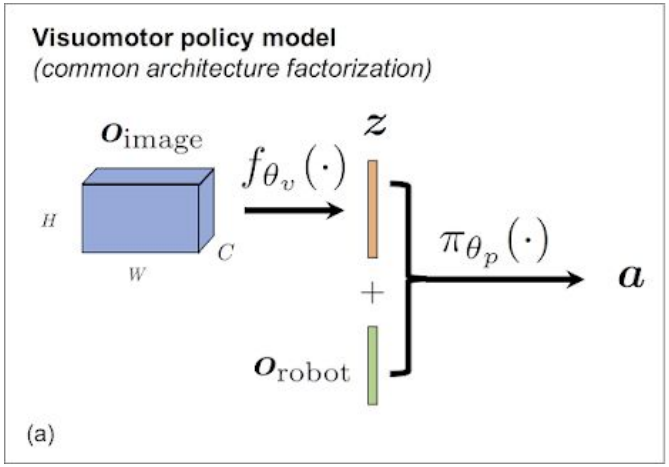
- Human assistance required during data collection*
- Mostly semi-rigid or rigid objects
- Limited semantic understanding - **R3M** (Nair et al.)

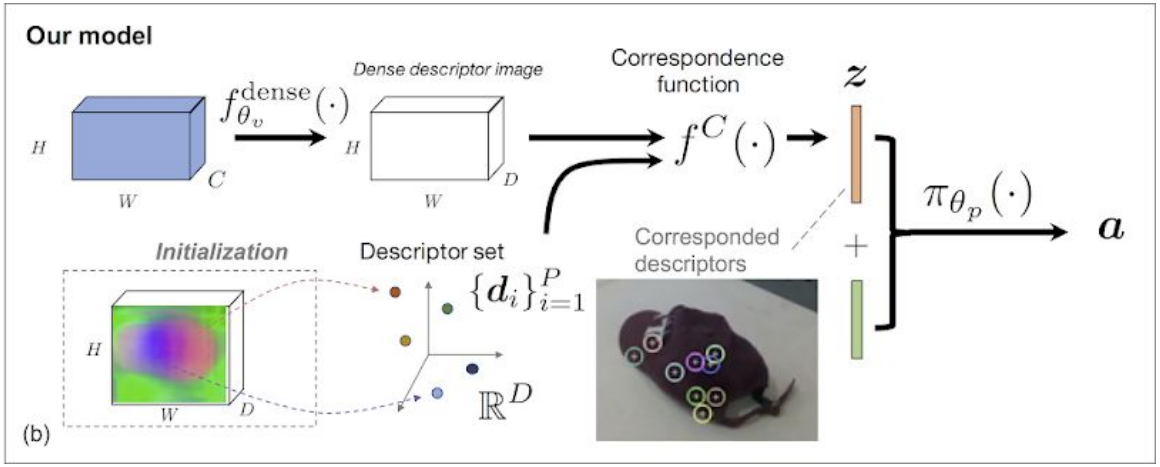
Questions

- How useful are the representations for learning visuomotor policies?
- What is the ideal representation for robotics?

Follow-up Work

Self-Supervised Correspondence in Visuomotor Policy Learning (Florence et al.)





Extended Readings

- Core
  - Schmidt, T., Newcombe, R., & Fox, D. **Self-supervised visual descriptor learning for dense correspondence.**
  - Florence, P., Manuelli, L., & Tedrake, R. **Self-supervised correspondence in visuomotor policy learning.**

- Further
  - Shridhar, M., Manuelli, L., & Fox, D. **Cliport: What and where pathways for robotic manipulation.**
  - For more see references!

Summary

- Learning representations is a classic problem in robotics
- The paper proposes learning dense correspondences for objects and give practical improvements on prior works
- Extend the method to multiple objects
- The learnt descriptors can
  - Generalize to new poses and orientations
  - Be **class general** or **instance specific**
  - Can be used for specifying robot grasps

References

[1] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation

[2] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images

[3] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates

[4] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network

[5] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling

[6] T. Schmidt, R. Newcombe, and D. Fox. Self-supervised visual descriptor learning for dense correspondence