Google AI

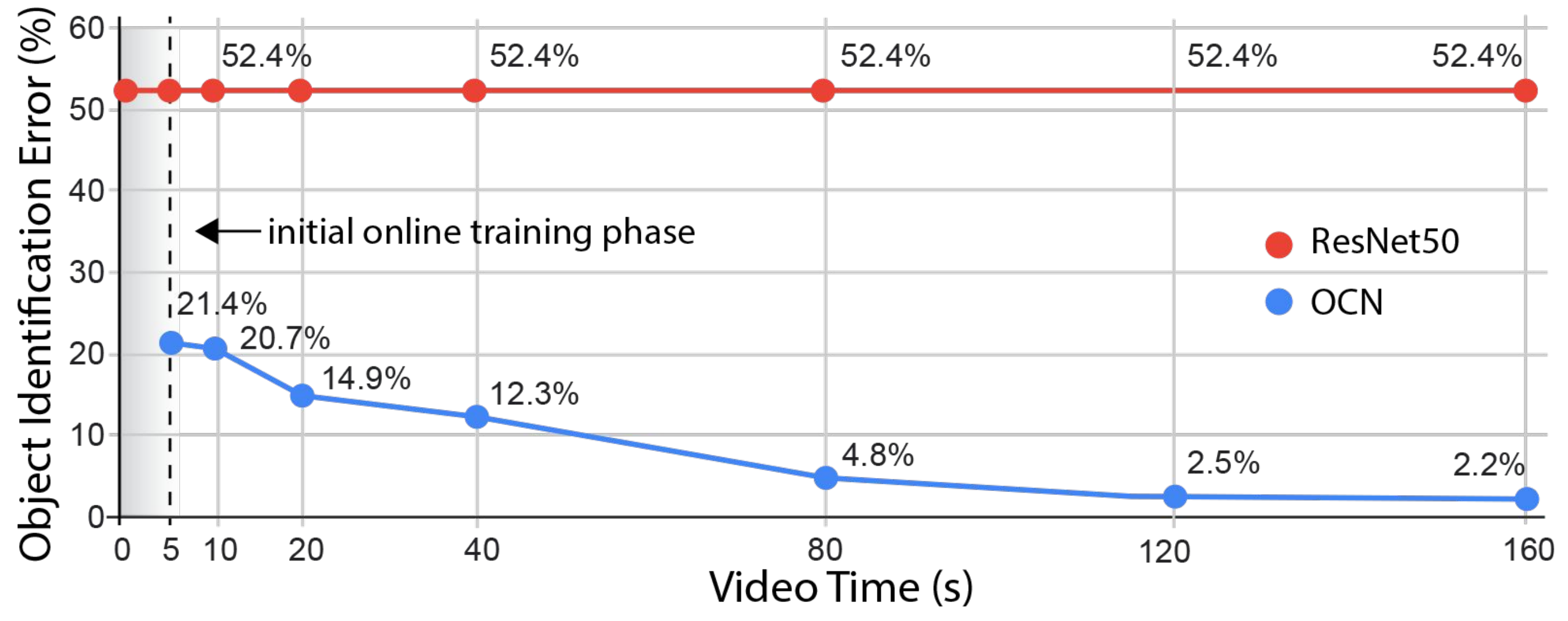
Online Object Representations with Contrastive Learning in Videos

Soren Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, Pierre Sermanet

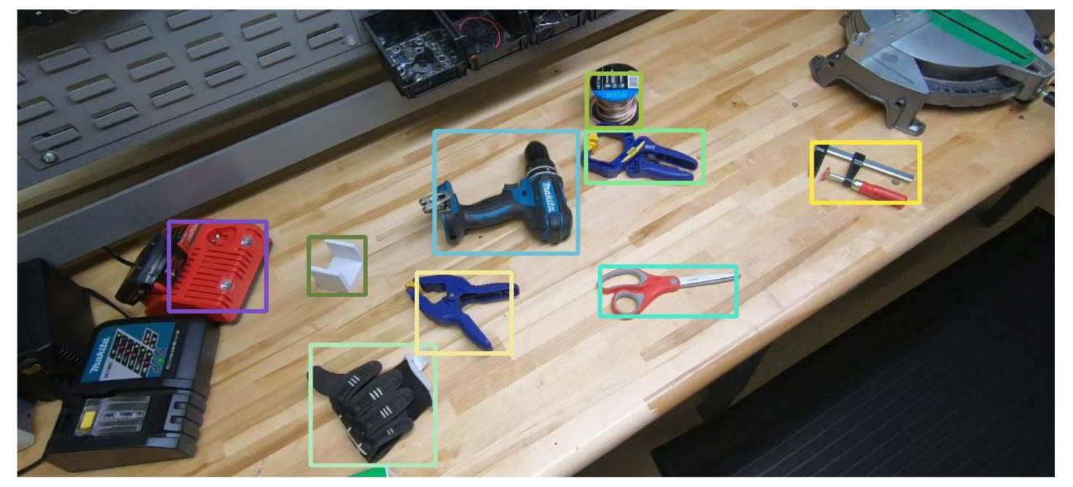
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Objective

- Self-teach to **discover** and disentangle **object attributes** from videos **without** using any **labels**.
- Use of **online adaptation**: the longer our online model looks at objects in a video, **the lower the object identification error.**
- Explore system **free of human supervision** for robotics applications. A **robot** collects its own data, trains on it, and then **identifies objects**.

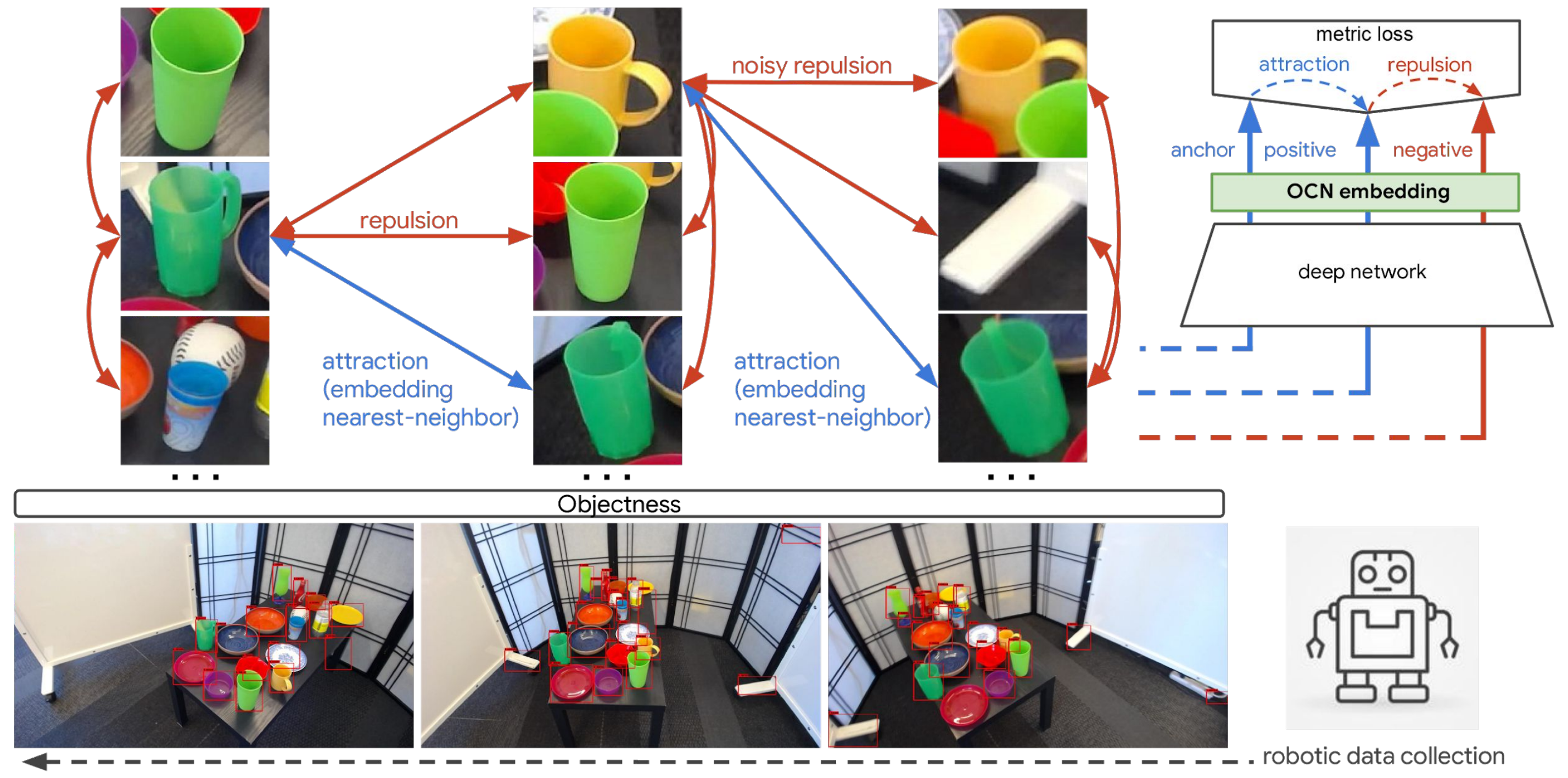


Frame n


Frame m

Approach

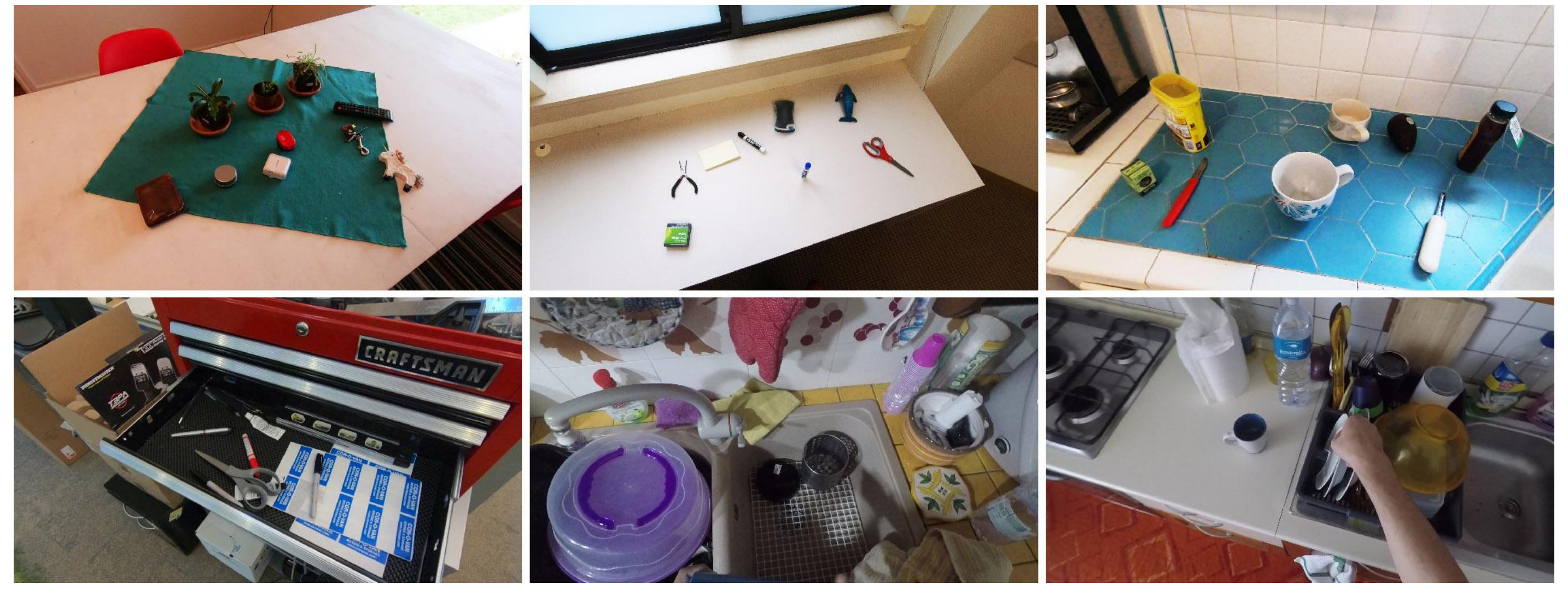
- Detect and **embed objects** to **extract** their **features.**
- Use **metric loss** to **contrast similar and dissimilar objects** in embedding space.
- **Observing objects** across different views **facilitates learning invariance** to scene-specific properties, such as scale, occlusion, lighting, or background.



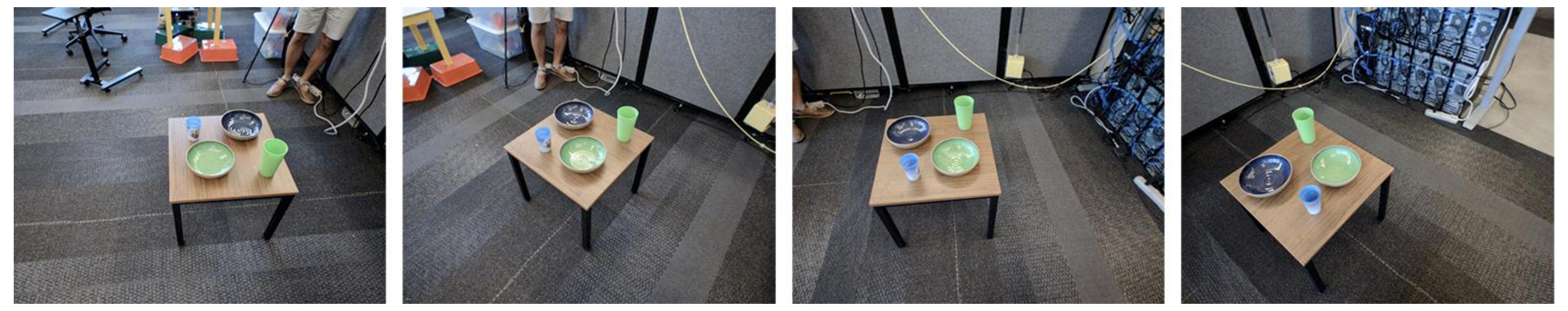
By **attracting** nearest neighbors in embedding space and **repulsing** others using **metric learning**, continuous **object representations** naturally emerge.

Datasets

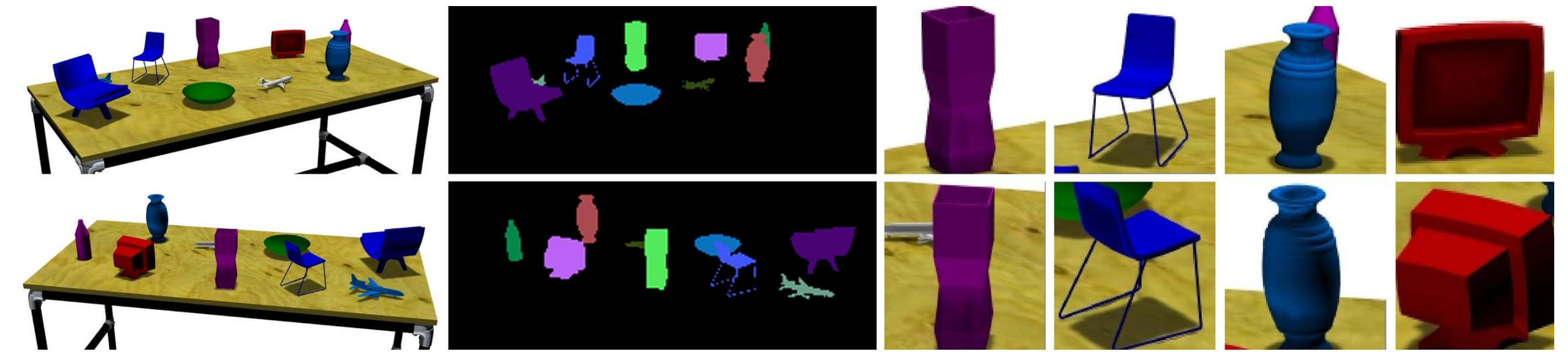
Real Data for Online Training (Complex Scenes, Epic Kitchens)



Automatic Real Data Collection with Robot

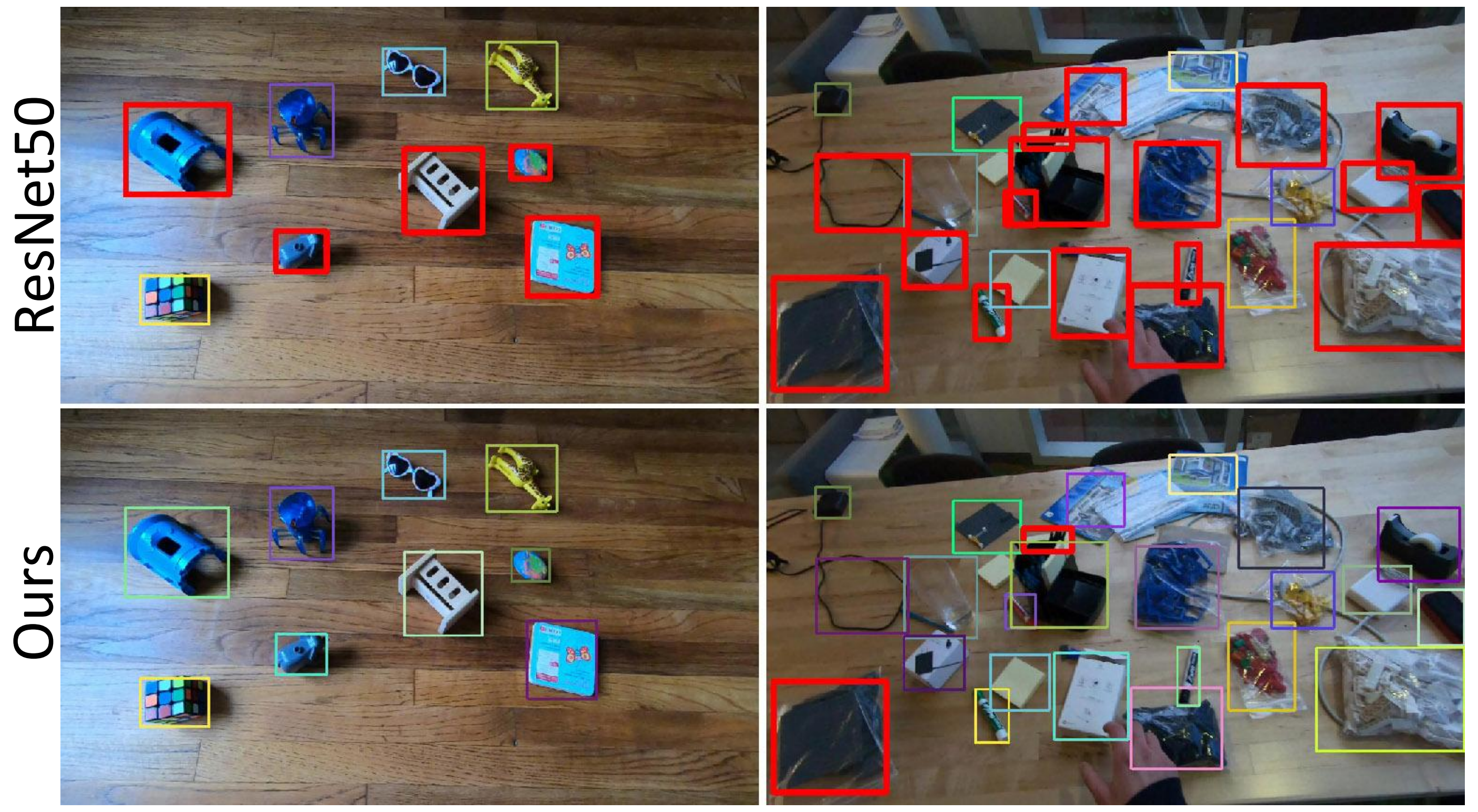


Synthetic Data for Evaluation

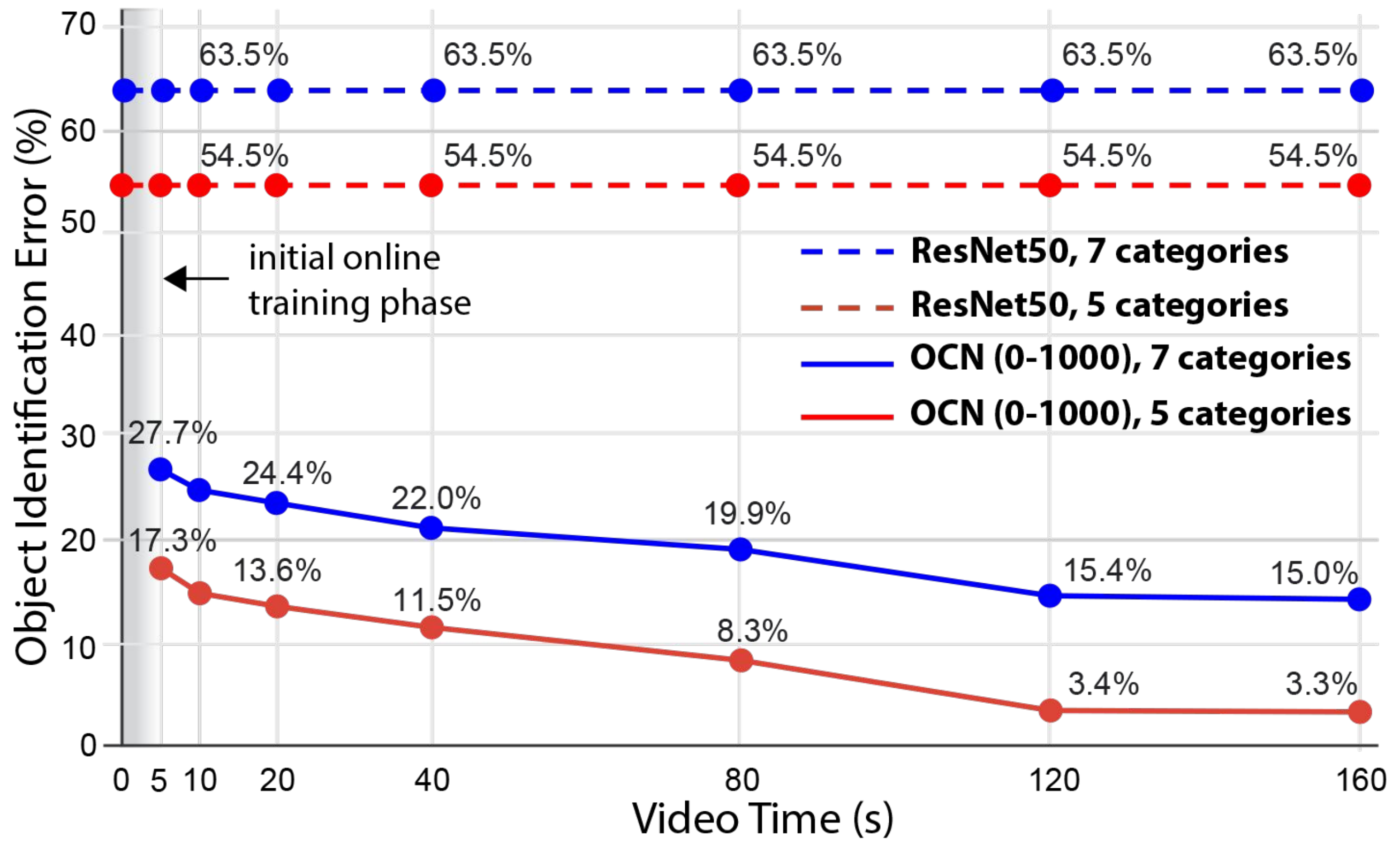


Experiments

Online Object Identification (red boxes indicate mismatches)



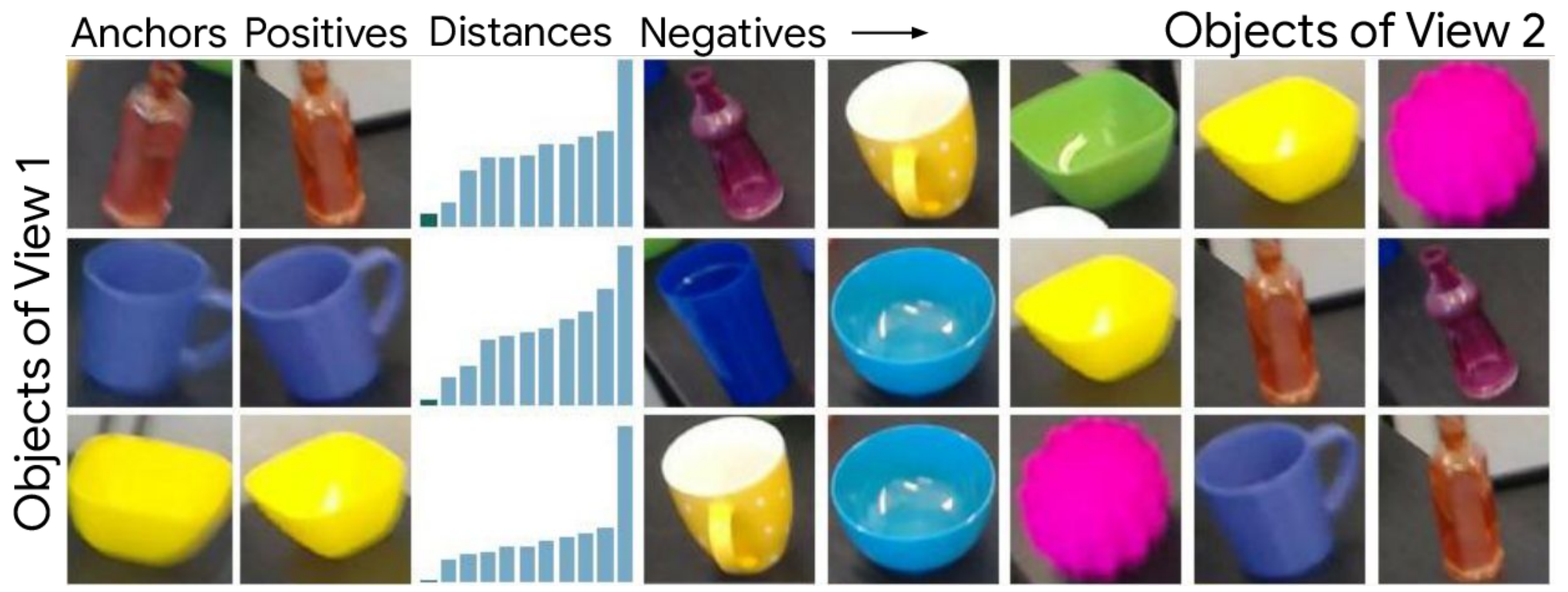
Self-supervised **online training** enables adapting to **unseen objects**, important for **robotic agents**.



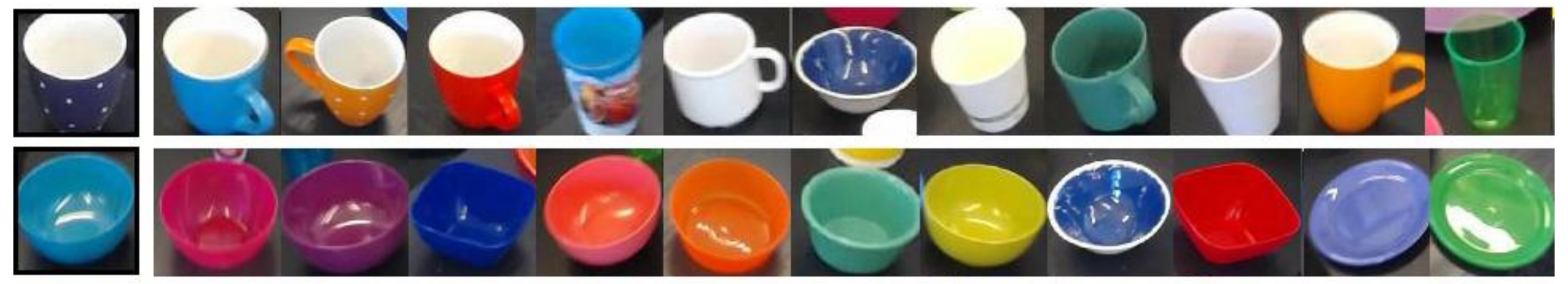
Object Attribute Classification: Comparison to Baselines

Method	Class (12) Attribute Error	Color (8) Attribute Error	Binary Attributes Error	Embedding Size
[BL] Softmax	2.98%	0.80%	7.18%	-
[BL] OCN sup (linear)	7.49%	3.01%	12.77%	32
[BL] OCN sup (NN)	9.59%	3.66%	12.75%	32
[ours] OCN unsup. (linear)	10.70%	5.84%	13.76%	24
[ours] OCN unsup. (NN)	12.35%	8.21%	13.75%	24
[BL] ResNet50 embed. (NN)	14.82%	64.01%	13.33%	2048
[BL] Random Chance	91.68%	87.50%	50.00%	-

View to View Correspondence (nearest neighbors, same scene)

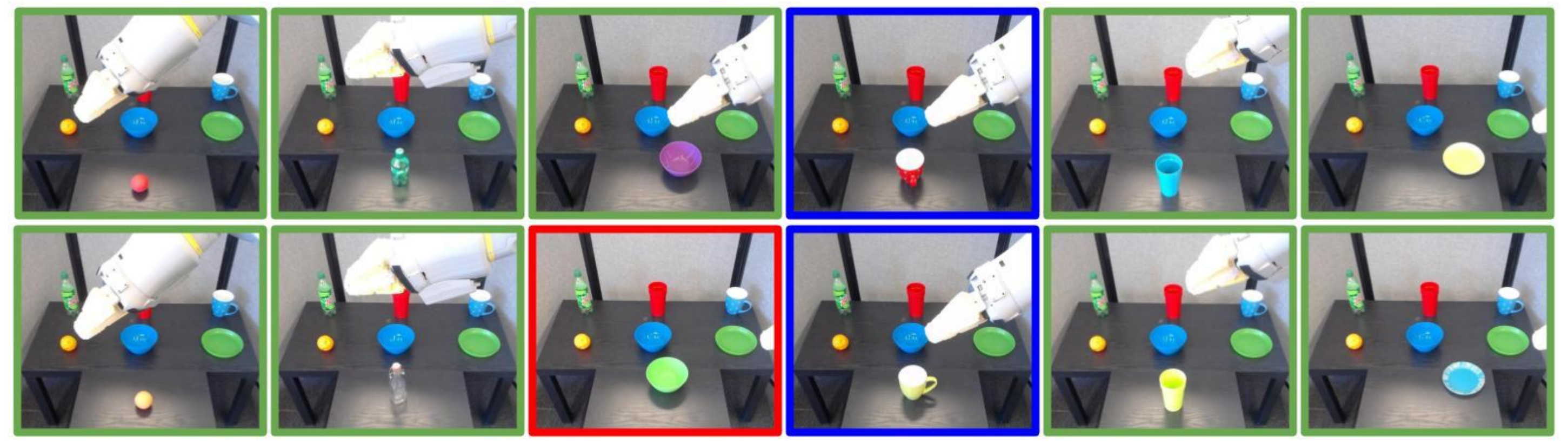


Feature Alignment (nearest neighbors, dataset)



Our approach allows to **organize objects** along their **visual** and **semantic** properties.

Robotic Pointing



Point at **object** that is **most similar** to the one shown.

Conclusion

Self-supervised online learning of **object representations**, particularly useful for **robotics** to increase robustness and adaptability to **unseen objects.**

Paper and Videos available here:
https://online-objects.github.io/