Learning Dense Object Descriptors from Multiple Views for Low-shot Category Generalization

Stefan Stojanov, Anh Thai, Zixuan Huang, James M. Rehg

Georgia Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

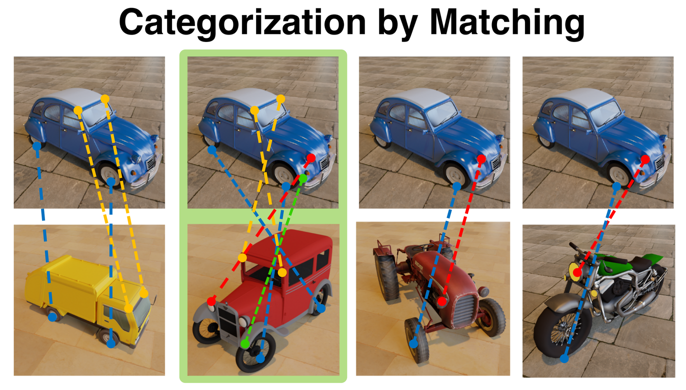
Code Available

Local Part Matching → Low-Shot Learning

Object Parts – Constituent Elements

- Consistently occur across instances
- Removal significantly alters object
- Mutually distinct
- Defined by non-accidental properties
  - Do not depend on viewpoint or object pose

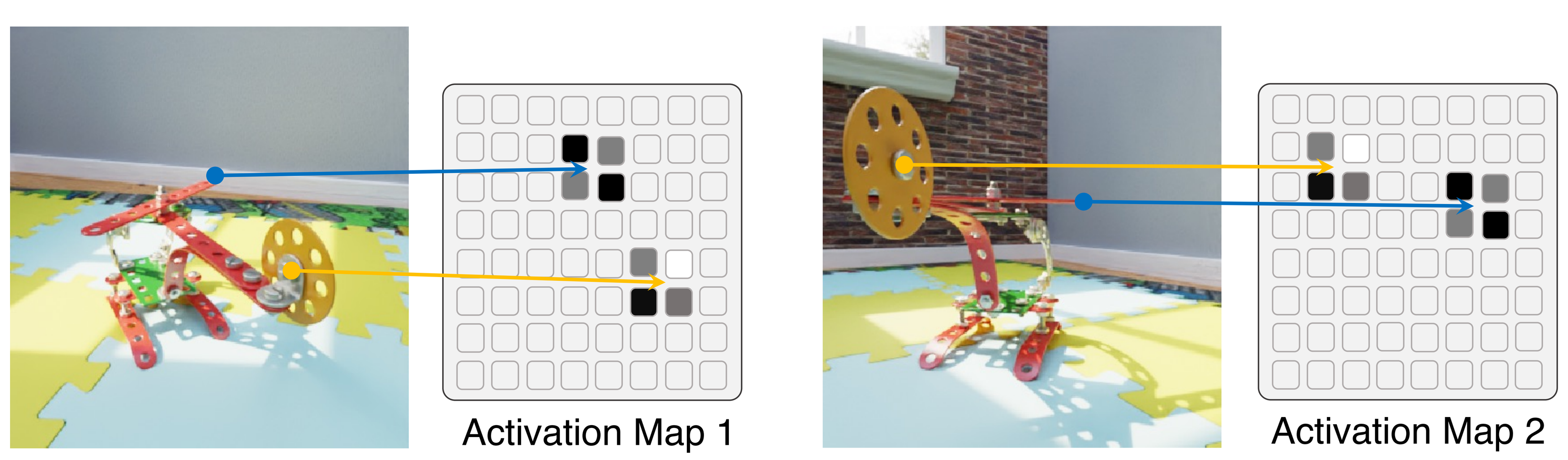
Categorization by Matching



most matches!

Key Idea: Learning to match local object parts across views
→ low-shot categorization by matching parts across instances

Part Descriptor Properties: Across Views



Non-accidental → invariant features

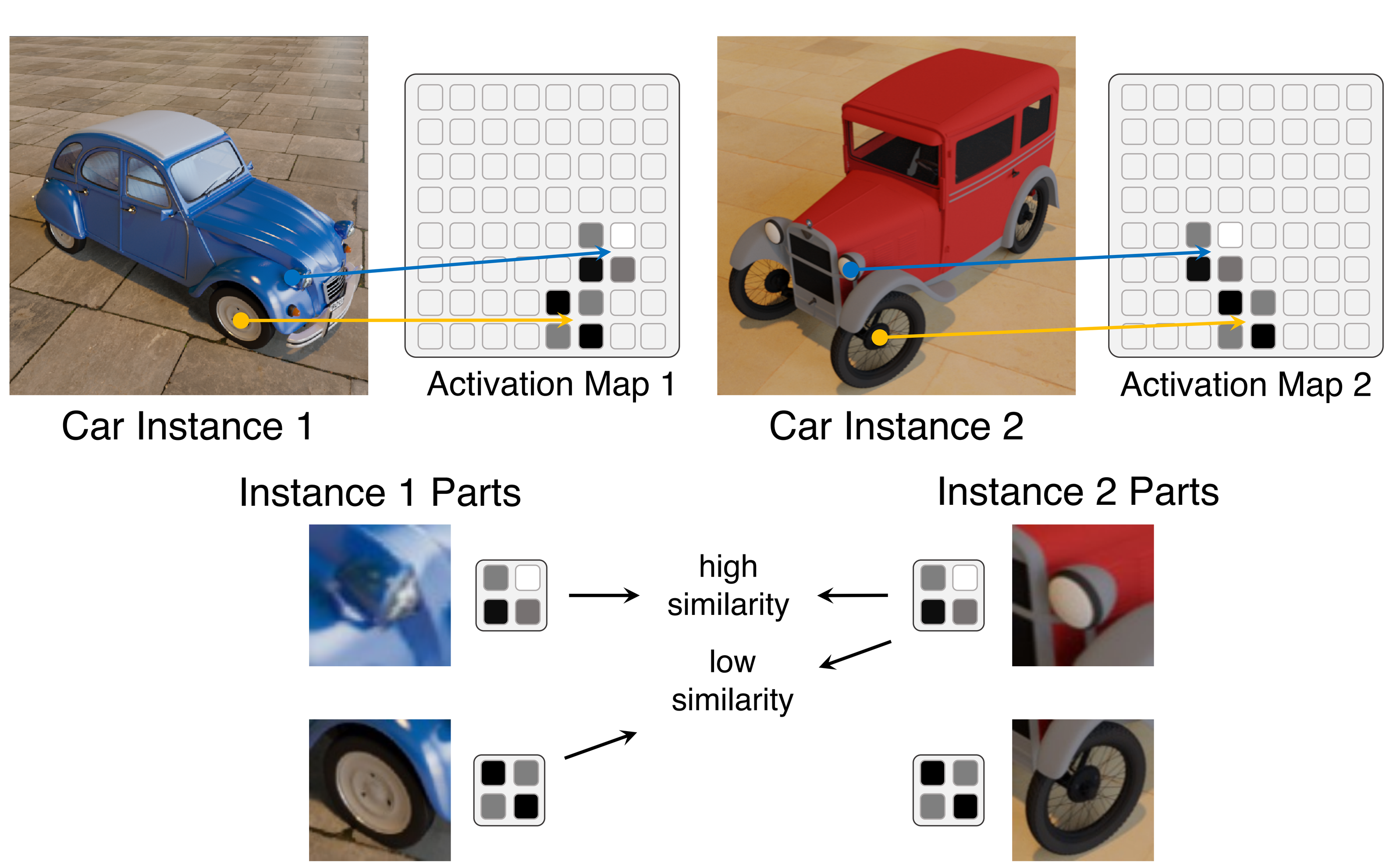
- Same features for whichever pixels the object part is projected to

Mutually Distinct

- Different parts encoded using different features

Desired properties can be learned from multiple views of objects!

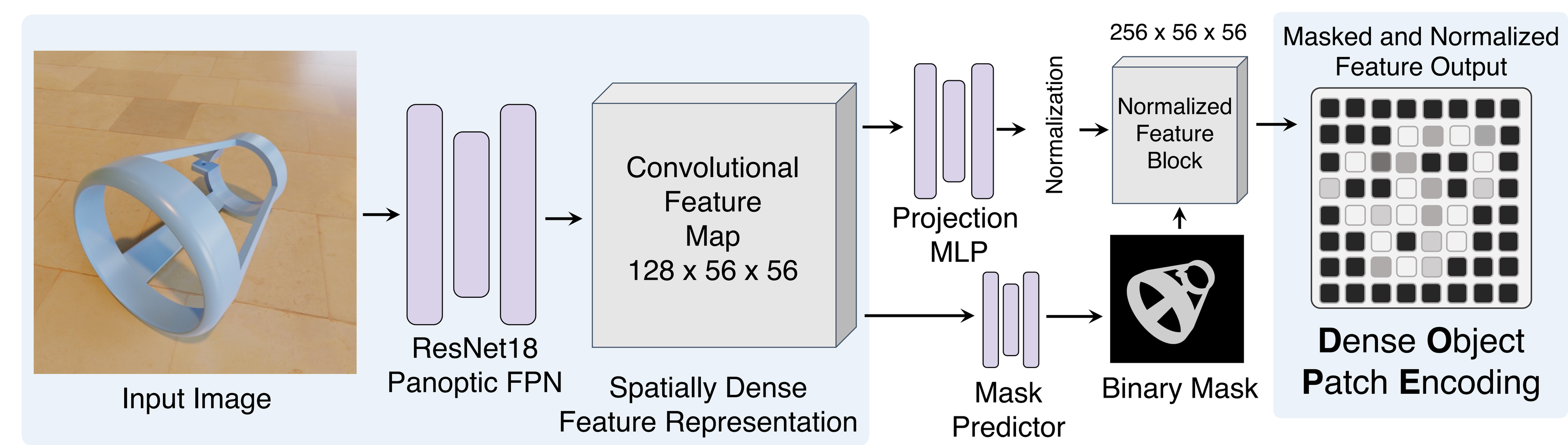
Part Descriptor Properties: Across Instances



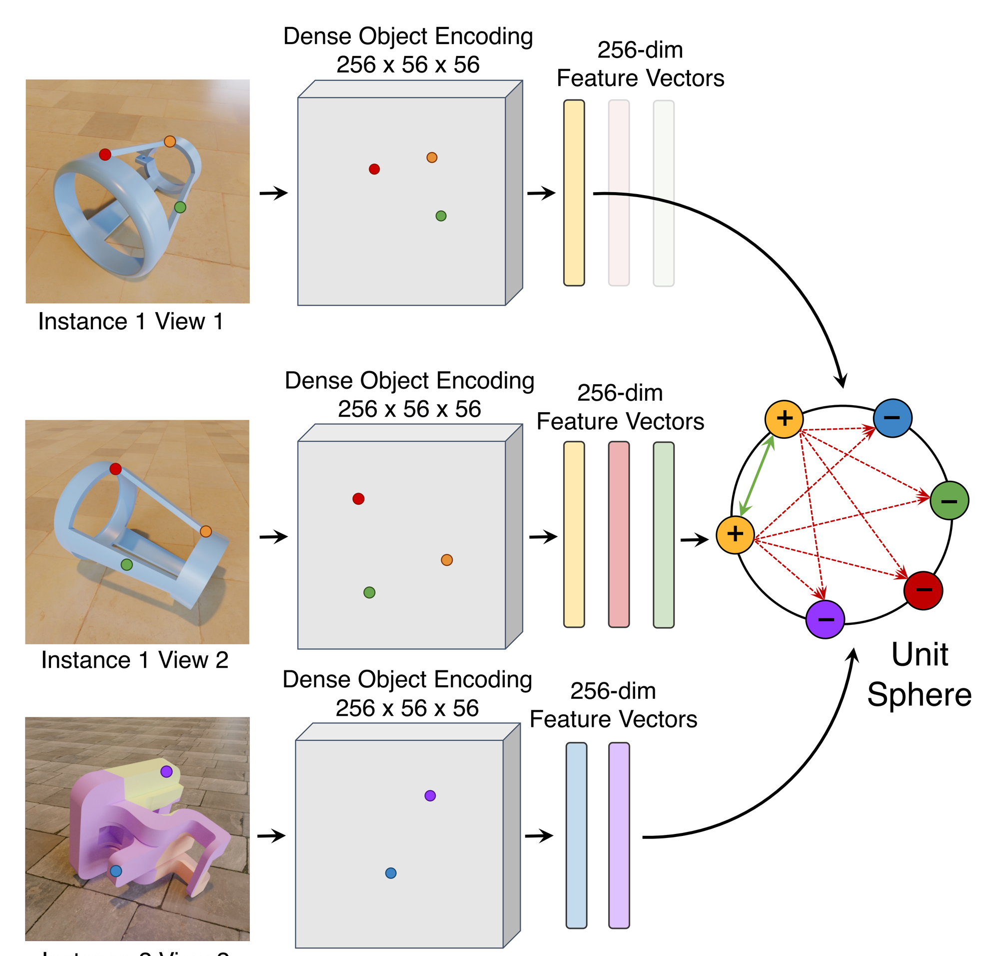
Hypothesis: Category-level descriptors can emerge by training across views of single instances

Multi-view self-supervised learning allows for low-shot object category recognition

Architecture

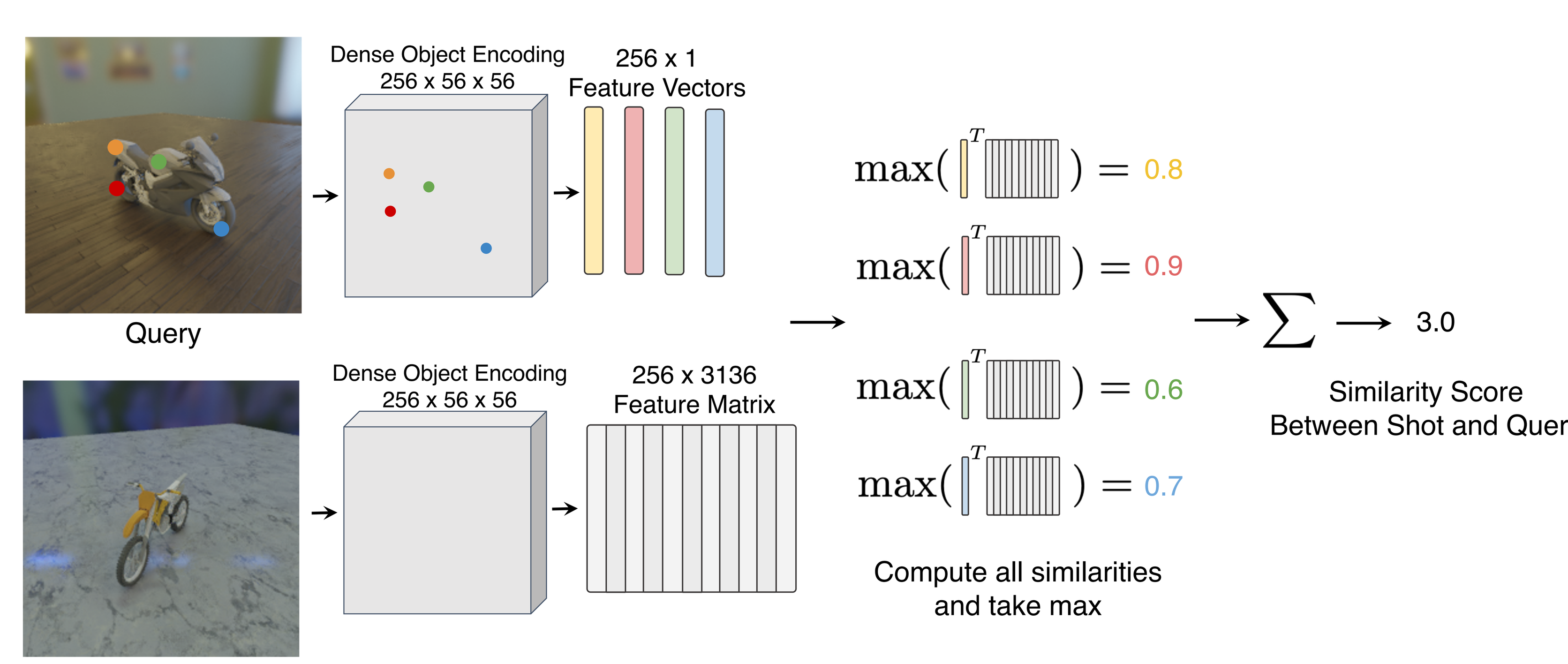


Multi-View Self Supervised Learning

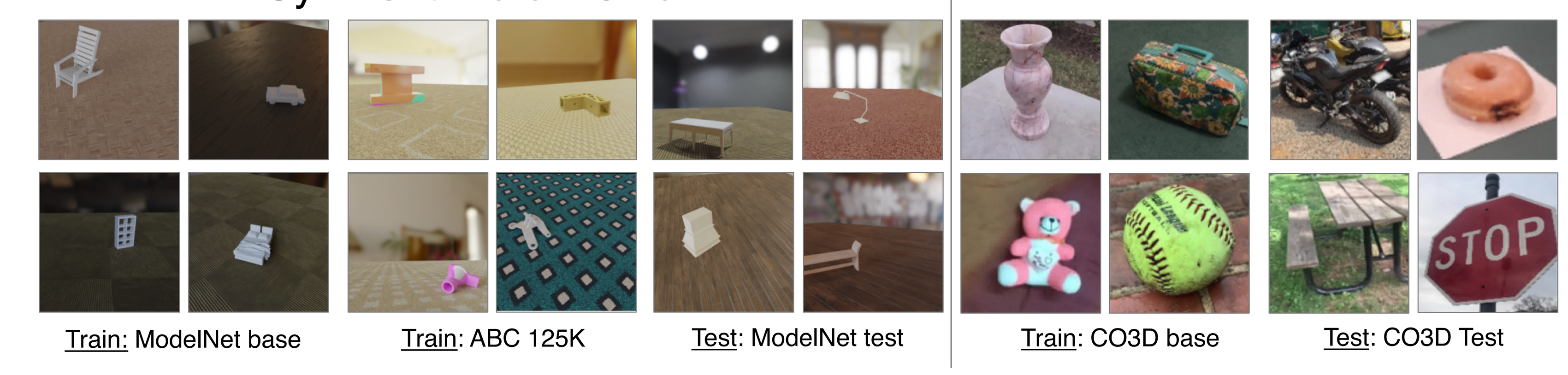


1. Sample points on the object surface using farthest point sampling
2. Find corresponding points in 2nd view using known viewpoint geometry
3. Extract local features based on point locations
4. Use corresponding positives and negatives for contrastive training using the InfoNCE loss

Low-Shot Learning by Matching



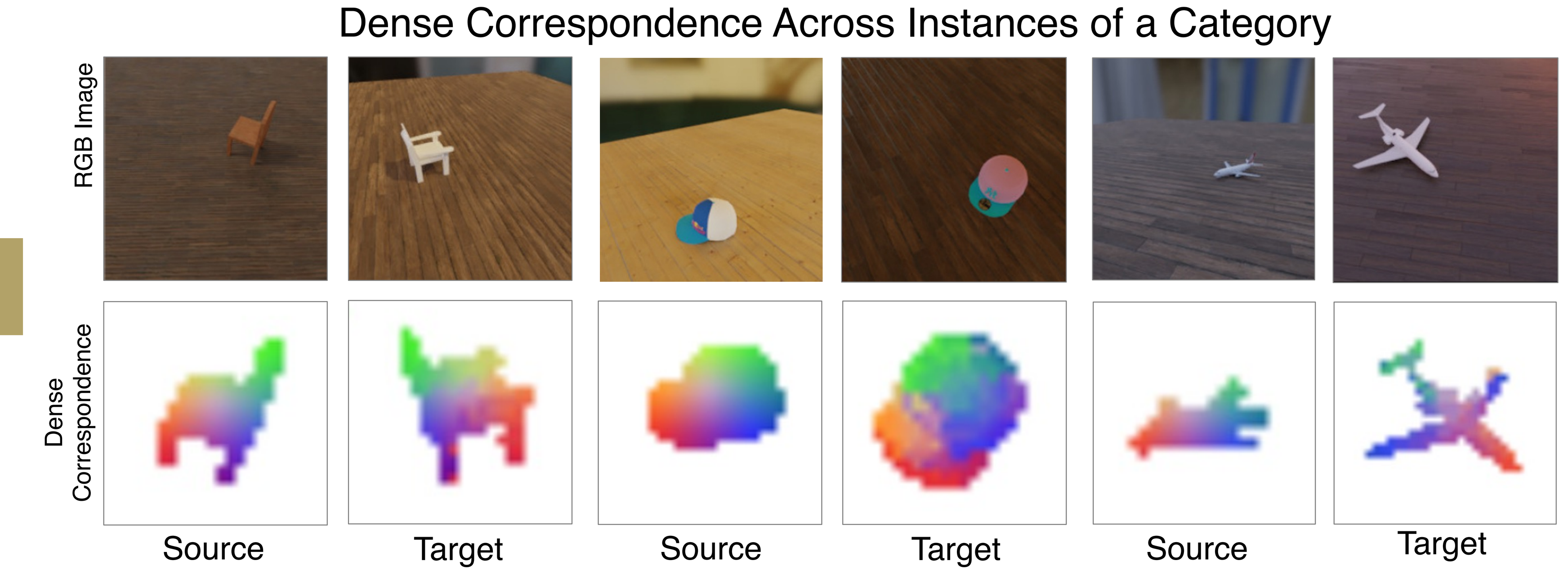
Evaluation Setting



In the synthetic domain correspondences are obtained from known cameras and viewpoints, on real video data like CO3D they are obtained with COLMAP

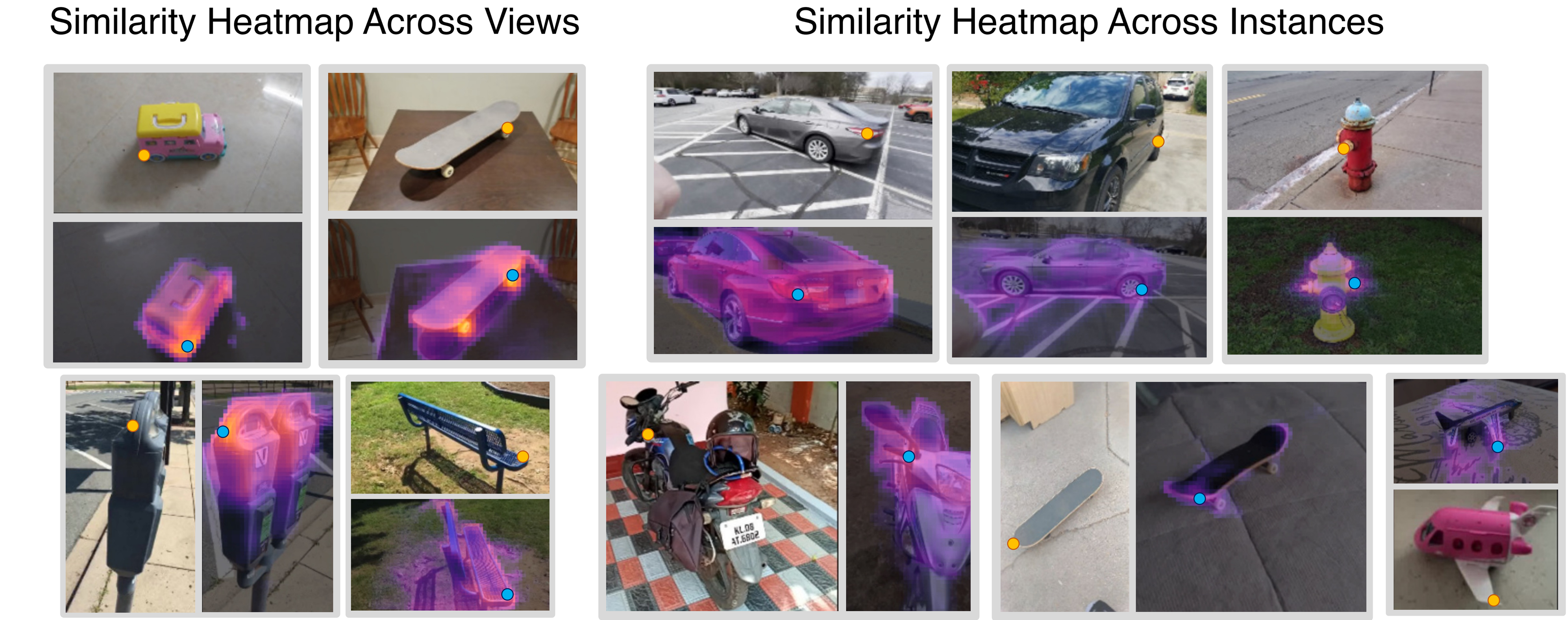
Qualitative Results

Dense Correspondence Across Instances of a Category



Similarity Heatmap Across Views

Similarity Heatmap Across Instances



Quantitative Results on ModelNet

		5-classes		10-classes	
		1-shot	5-shot	1-shot	5-shot
Supervised	SimpleShot	56.55 (± 0.42)	69.87 (± 0.32)	41.27 (± 0.24)	54.84 (± 0.18)
	RFS	57.31 (± 0.30)	73.77 (± 0.33)	42.22 (± 0.34)	59.97 (± 0.18)
	FEAT	57.46 (± 0.39)	71.73 (± 0.32)	41.72 (± 0.24)	57.84 (± 0.18)
	SupMoCo	55.32 (± 0.40)	71.82 (± 0.33)	39.87 (± 0.23)	57.15 (± 0.17)
Self-Supervised	VISPE	56.27 (± 0.44)	67.76 (± 0.35)	40.41 (± 0.26)	51.97 (± 0.18)
	VISPE++ SimSiam	53.83 (± 0.29)	68.75 (± 0.25)	39.84 (± 0.17)	54.34 (± 0.13)
	VISPE++ MoCoV2	57.05 (± 0.42)	71.81 (± 0.35)	43.23 (± 0.25)	58.68 (± 0.18)
	DOPE (Ours)	57.51 (± 0.44)	70.44 (± 0.36)	42.73 (± 0.26)	55.52 (± 0.19)
	VISPE++ SimSiam - ABC	60.24 (± 0.28)	**76.55 (± 0.22)**	47.02 (± 0.18)	**64.51 (± 0.13)**
	VISPE++ MoCoV2 - ABC	61.07 (± 0.41)	75.96 (± 0.32)	47.67 (± 0.25)	63.27 (± 0.18)
	DOPE (Ours)	**62.76 (± 0.43)**	**76.86 (± 0.31)**	**49.39 (± 0.26)**	**64.77 (± 0.18)**