Pose-Aware Self-Supervised Learning with Viewpoint Trajectory Regularization

Berkeley University of California · University of Michigan · Caltech · UC Davis

Jiayun (Peter) Wang   Yubei Chen   Stella X. Yu

Motivation

- Car heading towards the camera
- At danger!

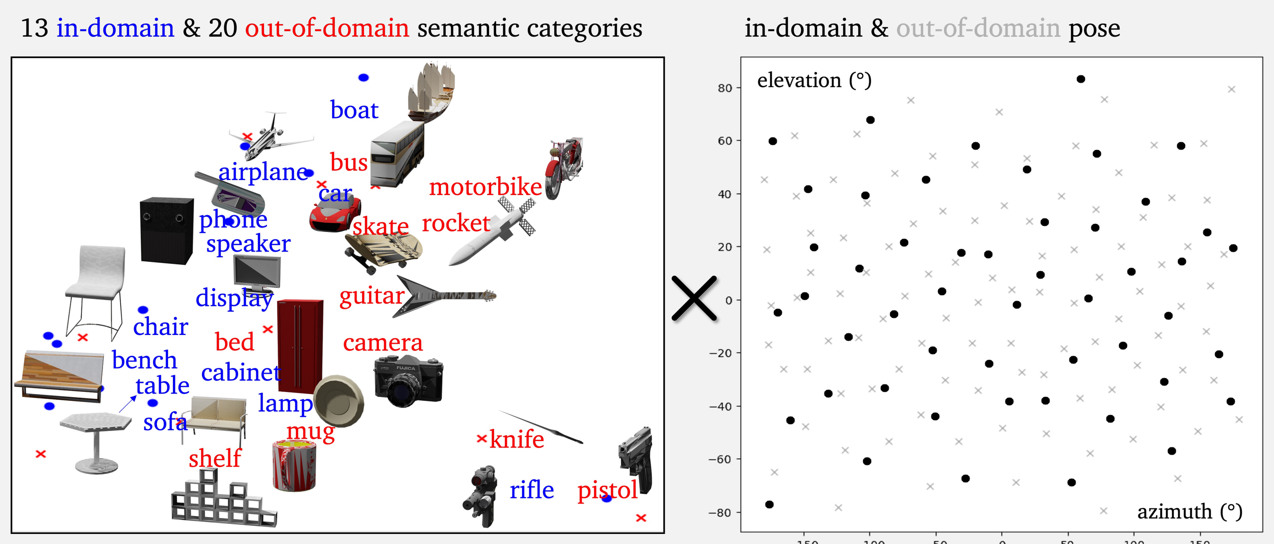
- Car heading away from the camera
- No danger

Recognition needs to understand two aspects:

- *What* is the object
- *How* is it presented

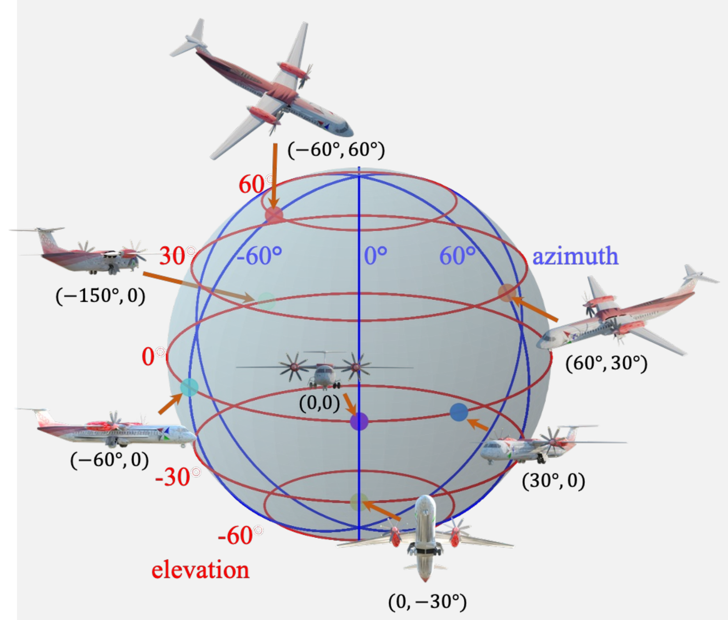
Benchmark

Data



Rendered images of objects of different categories and poses.
In-domain and out-of-domain splits for evaluating generalization.

Evaluation metrics



1) Semantic classification
2) Absolute pose
3) Relative pose
- Canonical pose free
- Category agnostic
- Generalize to open categories

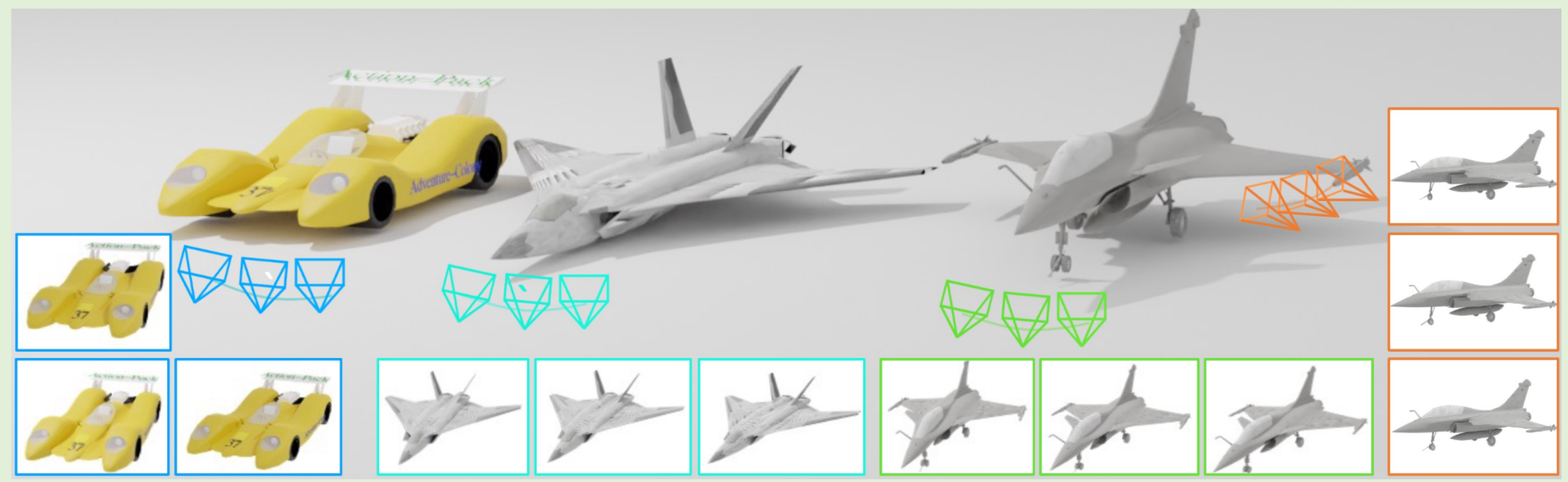
Unsupervised Learning for Both Object Semantics and Pose

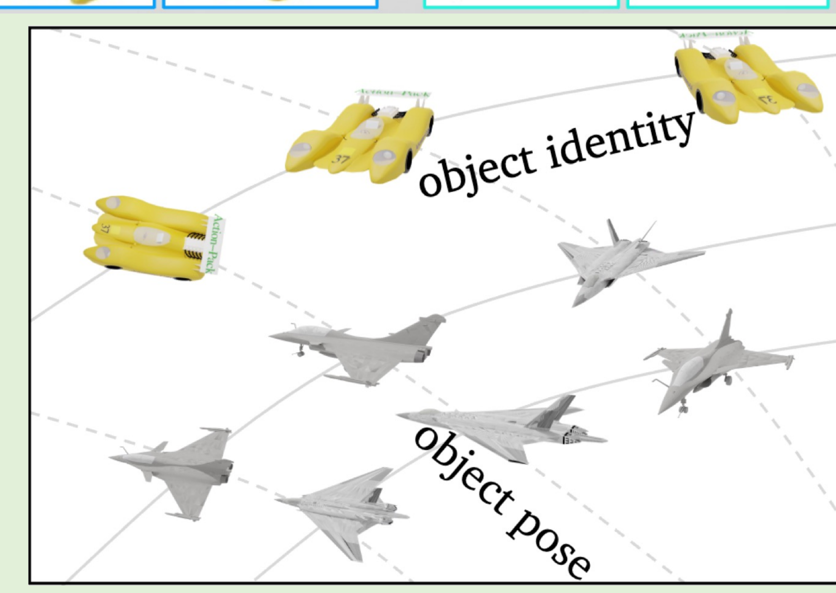
Goal: learn what is the object (semantics) and how it presented (pose)

Scenario: a robot moves around in the environment

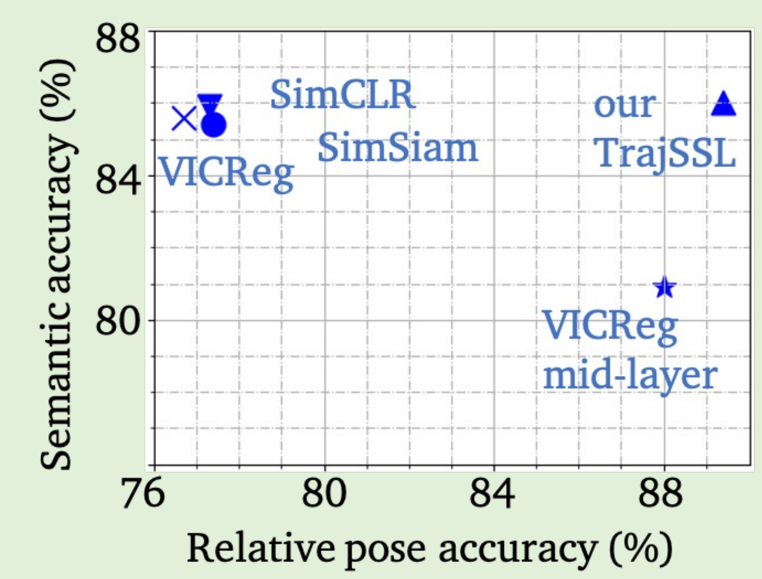
A natural **data acquisition** scheme:
- **No** labels
- **Adjacent** images of the same object from a smooth viewpoint trajectory

Training data: Image triplets with small pose changes; No semantic or pose labels.





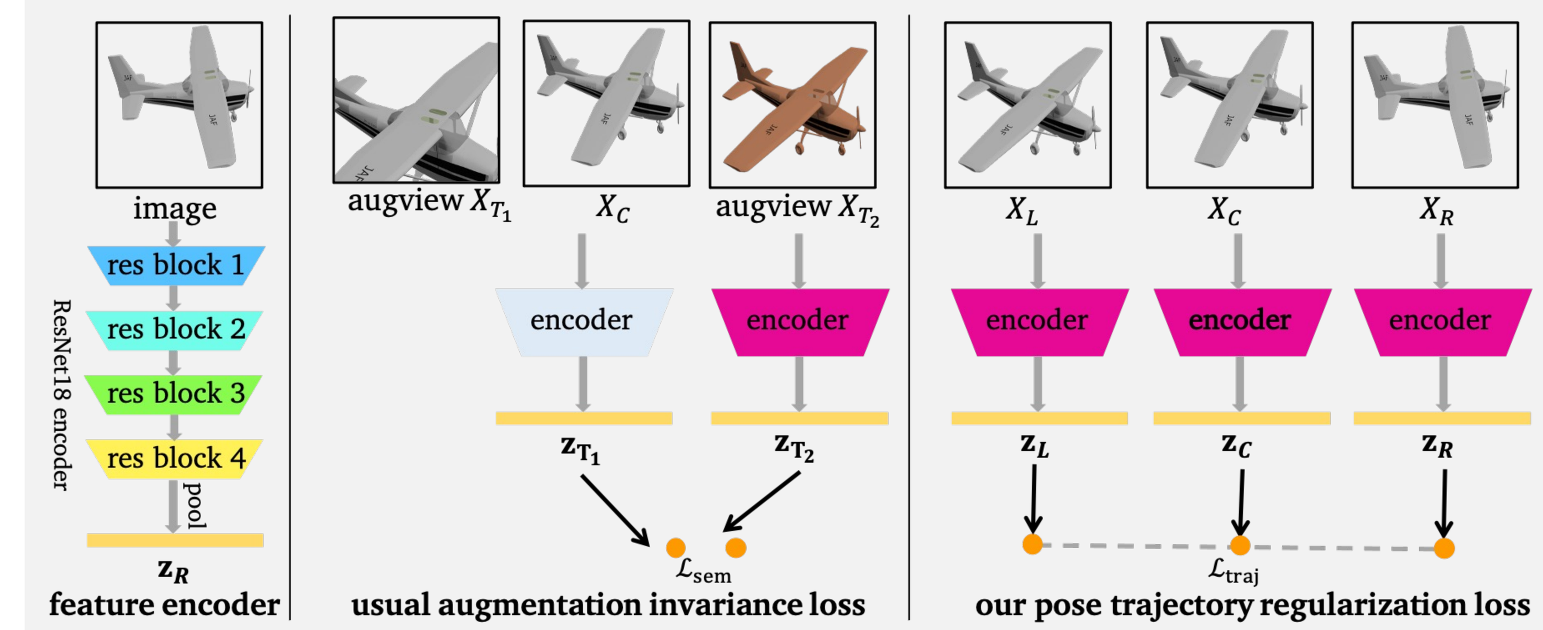
Emergent representation with disentangled semantics and pose



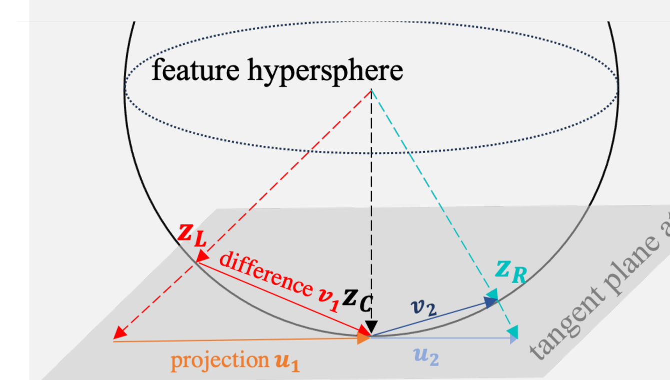
Gain in both semantic and pose accuracy.

Methods

Stage 1: Self-supervised representation learning



Our encoder produce embeddings for a triplet of images $\{X_L, X_C, X_R\}$ from a sequence with respective poses $\{p_L, p_C, p_R\}$ forming a trajectory, where pose changes are subtle.
Two unsupervised losses imposed on the embeddings, $\mathcal{L}_{sem}$ & $\mathcal{L}_{traj}$.
$\mathcal{L}_{sem}$ is an invariant loss (e.g. VICReg).

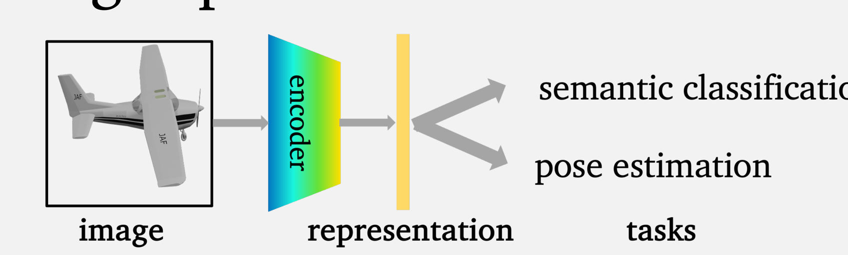


Trajectory Loss
Viewpoint trajectory regularization makes 3 embeddings form a line:

$$\mathcal{L}_{traj}(\mathbf{z_L}, \mathbf{z_C}, \mathbf{z_R}) = -\frac{\mathbf{u_1} \cdot \mathbf{u_2}}{\|\mathbf{u_1}\|\|\mathbf{u_2}\|}$$

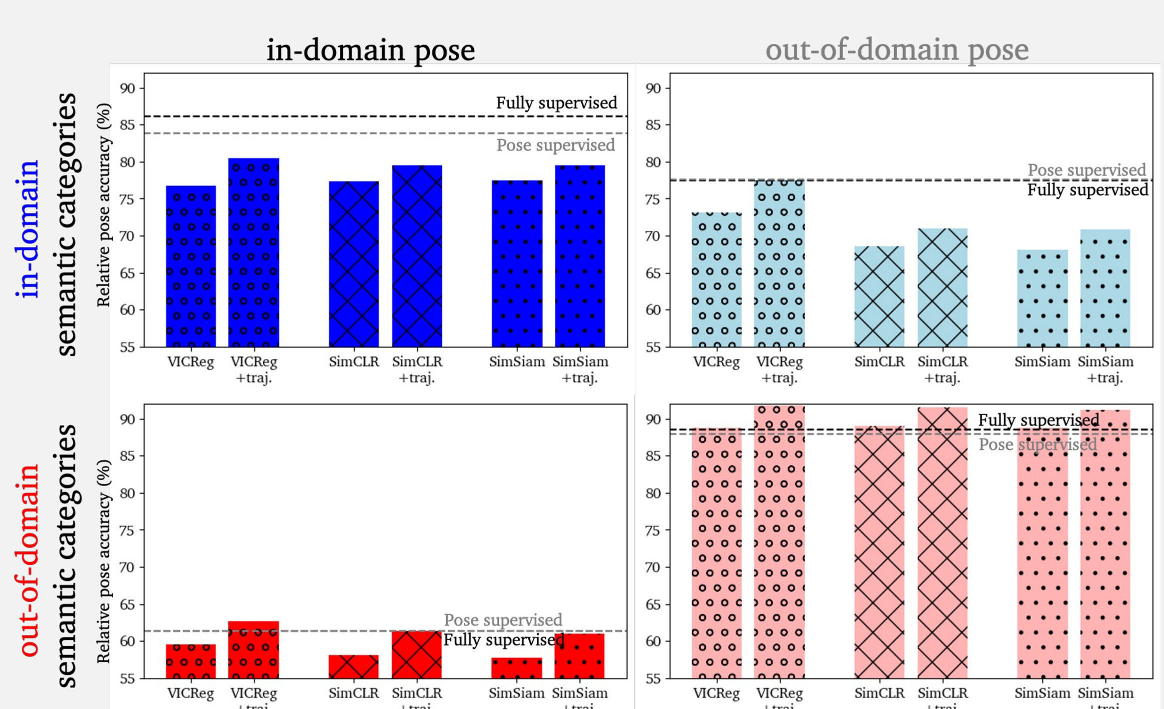
Final loss is a combination: $\mathcal{L} = \mathcal{L}_{sem} + \mathcal{L}_{traj}(\mathbf{z_L}, \mathbf{z_C}, \mathbf{z_R})$

Stage 2: Probing representation to downstream tasks



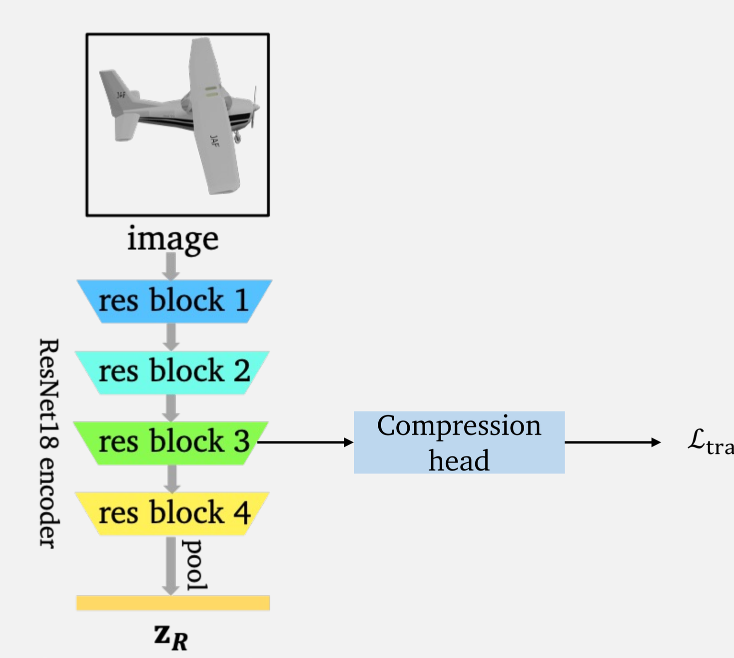
Results

Generalizability: Trajectory regularization improves relative pose accuracy for in and out-of-domain data.

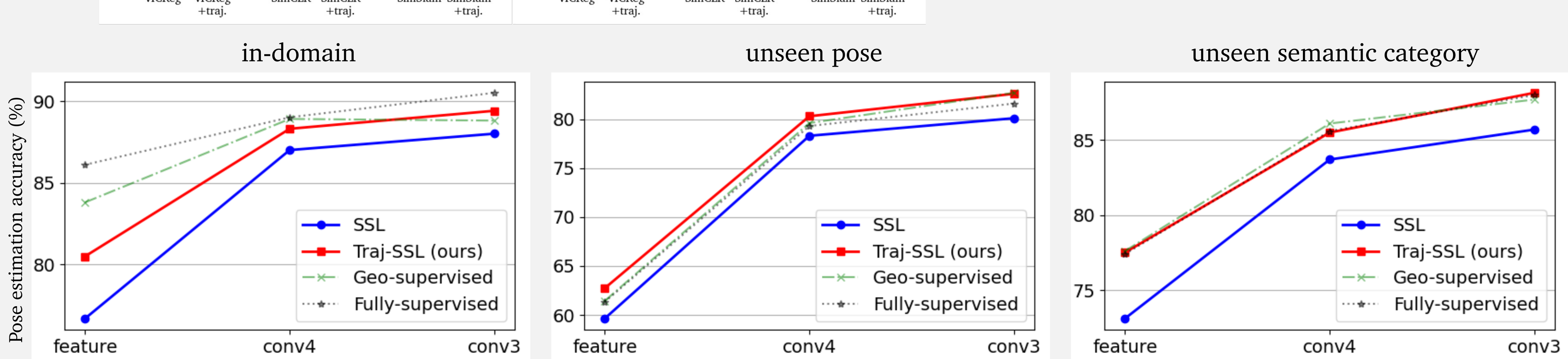


Real Data: Trajectory regularization improves retrievals with similar pose and appearance.

Compressing mid-layer representation up to 32x gives small accuracy loss.



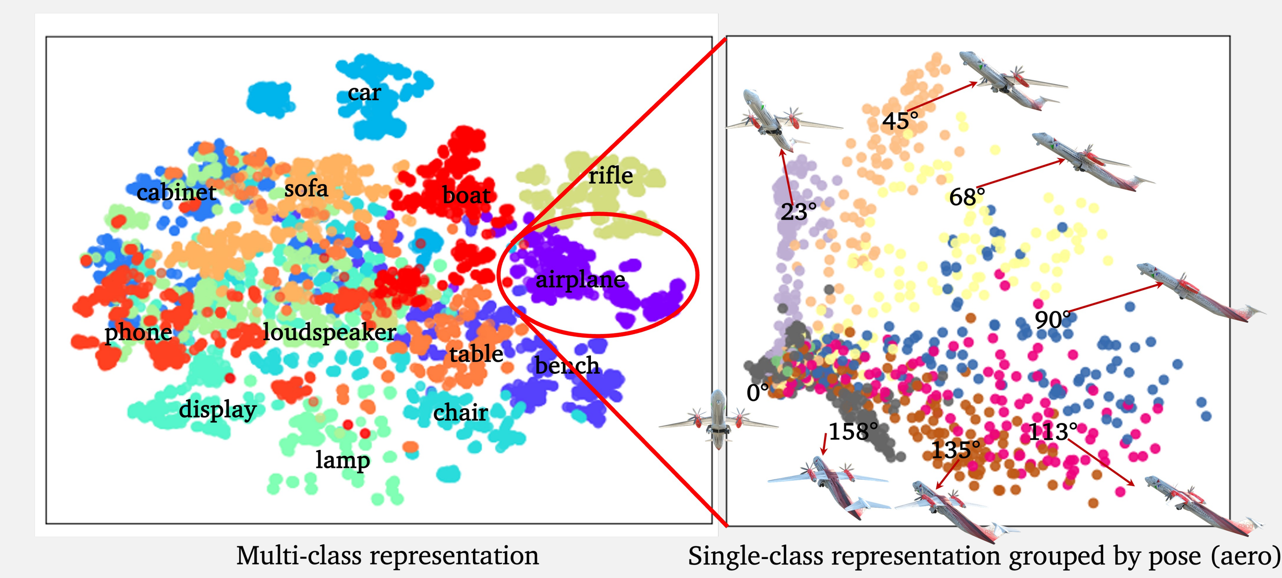
embedding	# dim	abs. pose acc. (%)	rel. pose acc. (%)
conv3	16,384	92.5	87.8
compressed conv3	512	91.4 (↓1.1)	82.4 (↓5.4)
conv4	8,192	91.9	85.2
compressed conv4	512	90.8 (↓1.1)	81.2 (↓4.0)
feature	512	87.8	77.5



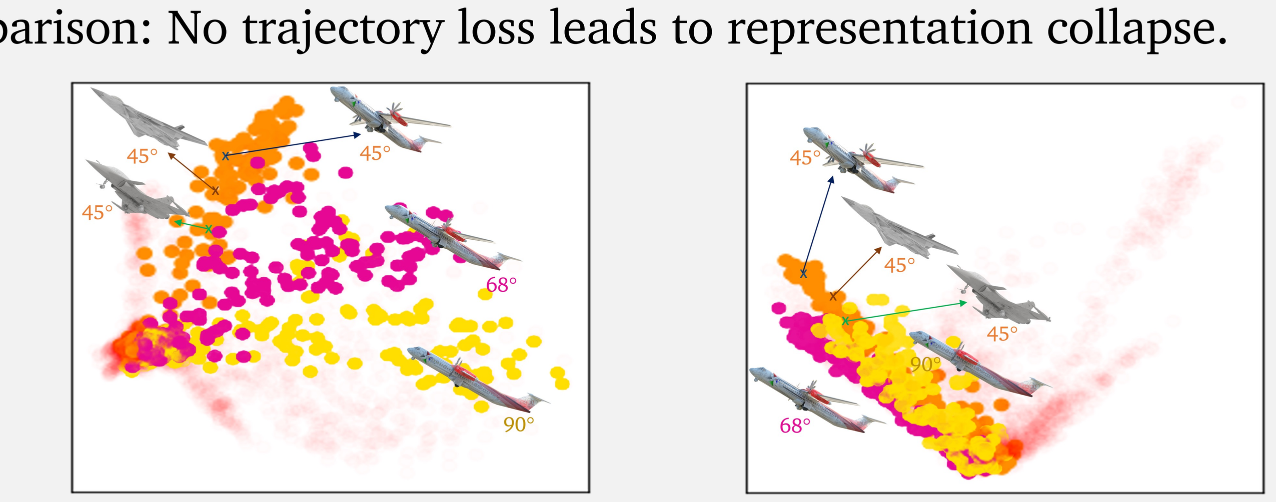
Mid-Layer Representation improves pose accuracy for in and out-of-domain data.

Representation Visualization

The joint semantic-pose embedding: Images are clustered by semantics; within each semantic cluster, images form mini-cluster by pose.



Comparison: No trajectory loss leads to representation collapse.



w/ trajectory loss   w/o trajectory loss (VICReg)