Team 06 ADLR

Self-supervised Learning for Robot Grasping

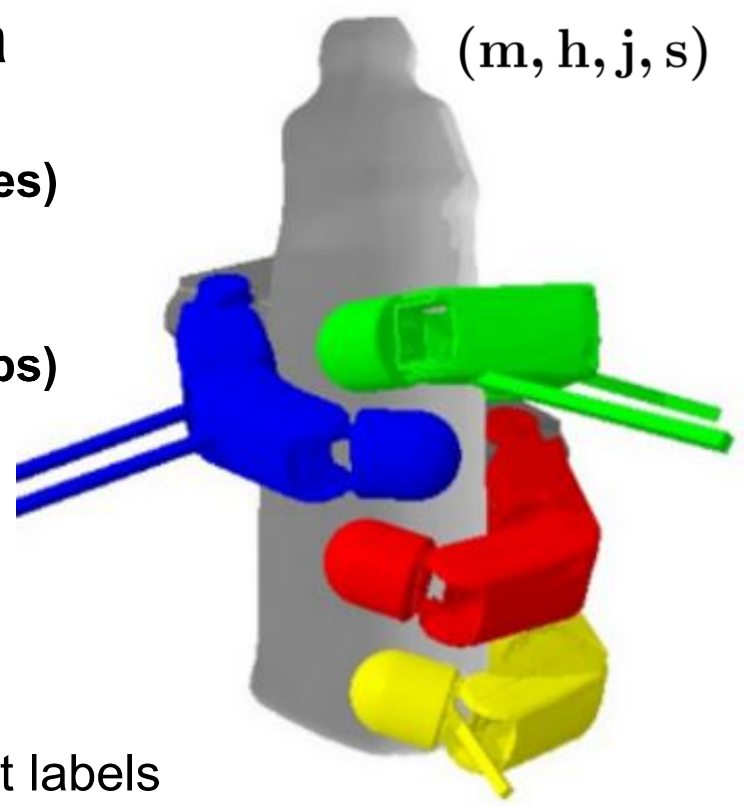
Jed Guzelkabaagac Boris Petrović
Technical University of Munich
{jed.guzelkabaagac, boris.petrovic}@tum.de

"Can we leverage unlabeled 3D data to make grasping more accurate and data-efficient?"

Data Scarcity: Labelled grasps expensive and slow to collect
Poor Generalization: Supervised-only models overfit data biases
Abundant Shape Data: Incorporate priors via unlabeled ShapeNet

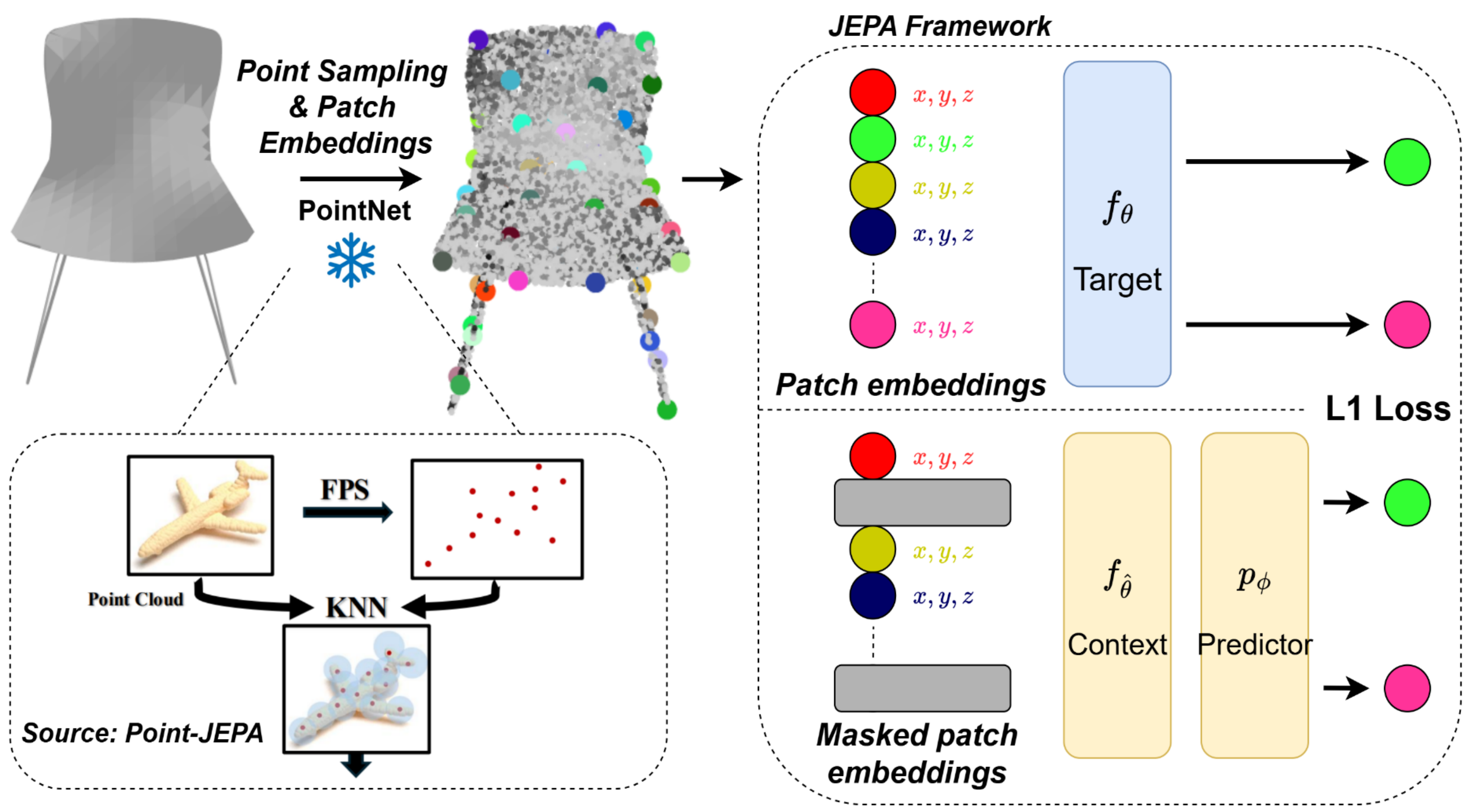
DLR-Hand II Grasping Data

m : **mesh (ShapeNet)** **(1.5k Meshes)**
$h \in \mathbb{R}^7$ **hand pose**
$j \in \mathbb{R}^{12}$ **joints pose**
$s \in \mathbb{R}$ **grasp score**
(373k Grasps)



Point-JEPA Pretraining

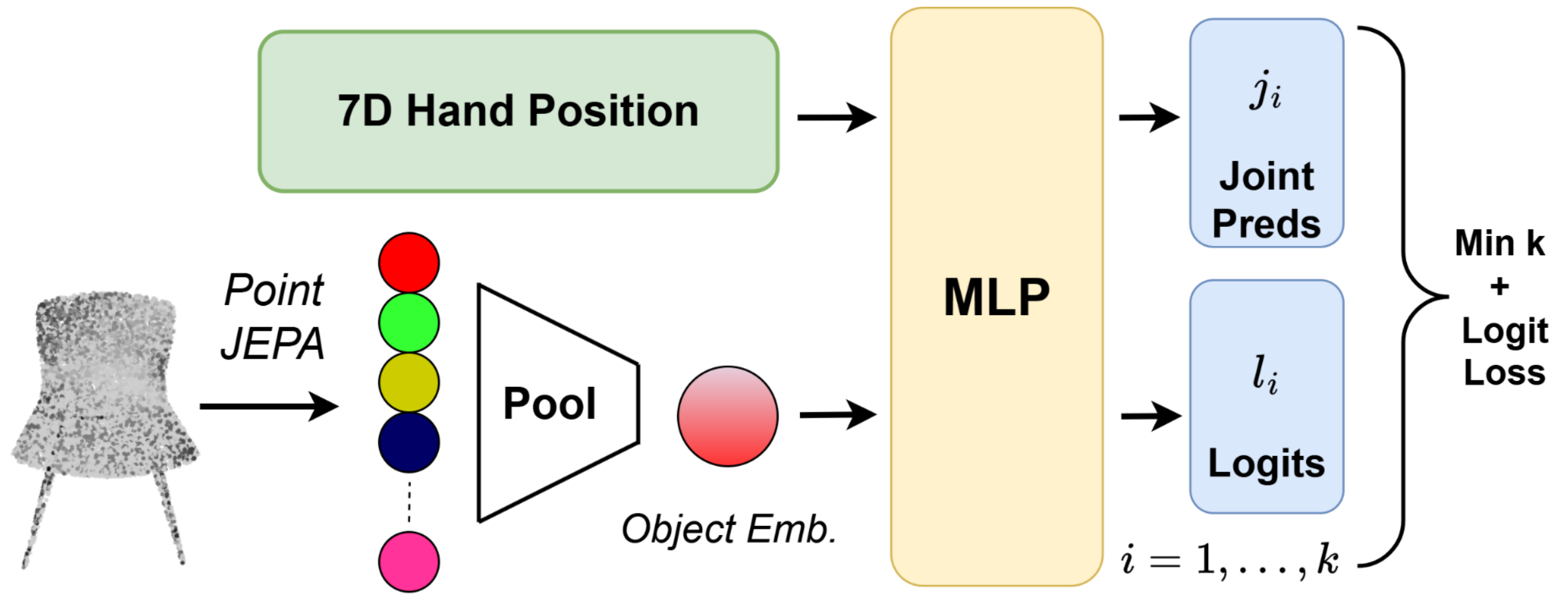
- Predicting masked parts of input
- Yields strong representations without labels
- Context-aware transformer for point cloud data



Objectives

- Do JEPA embeddings improve robot grasping?
- Can pretraining boost data efficiency in low-data regimes?
- Does it speed up learning or improve generalization?

Full Supervised Pipeline



Min-over-k Regression + Logit Loss

- **Multimodal grasps** – multiple valid solutions per object.
- K = 1 collapses to mean, loses diversity.
- Multi-head outputs + selection term preserve diversity.
- Logit allows utility at generation – output most confident.

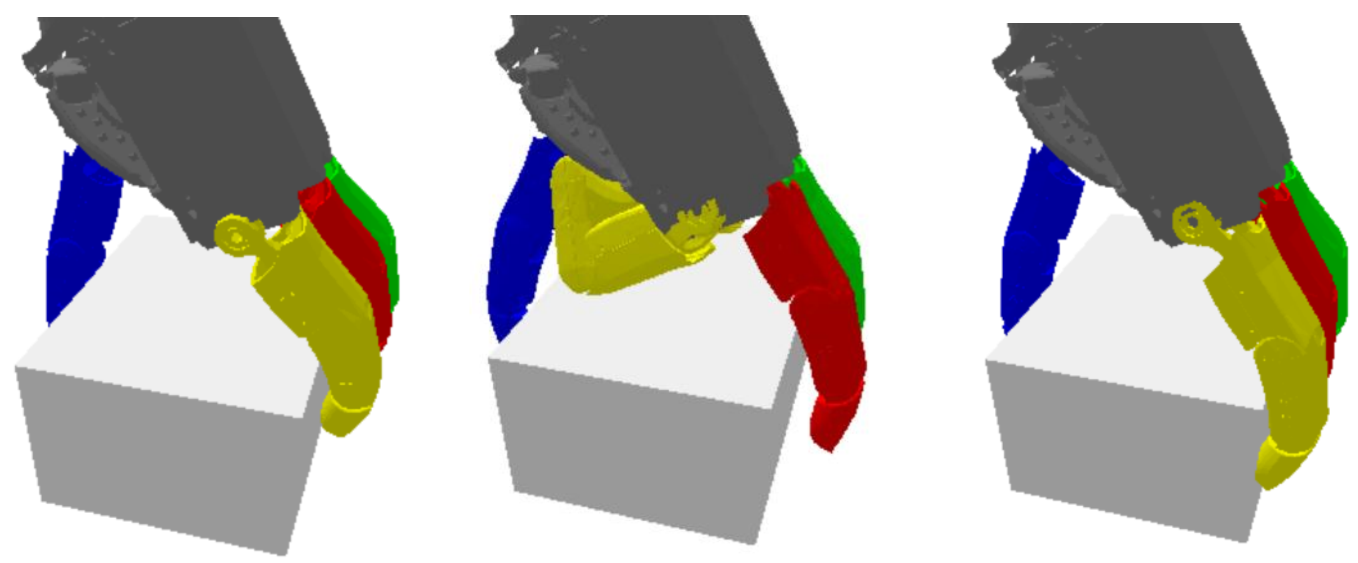
$$k^* = \arg\min_k \|\hat{\mathbf{j}}_k - \mathbf{j}\|^2$$

Index of best grasp

$$L = \|\hat{\mathbf{j}}_{k^*} - \mathbf{j}\|^2 + \alpha \cdot \mathrm{CE}(\ell, k^*)$$

Min L2 and learn best grasp output

Prediction 1 **Prediction 2** **Ground truth**



In this example, **collapsing to the mean** grasp would cause collisions.

Training Details & Ablations

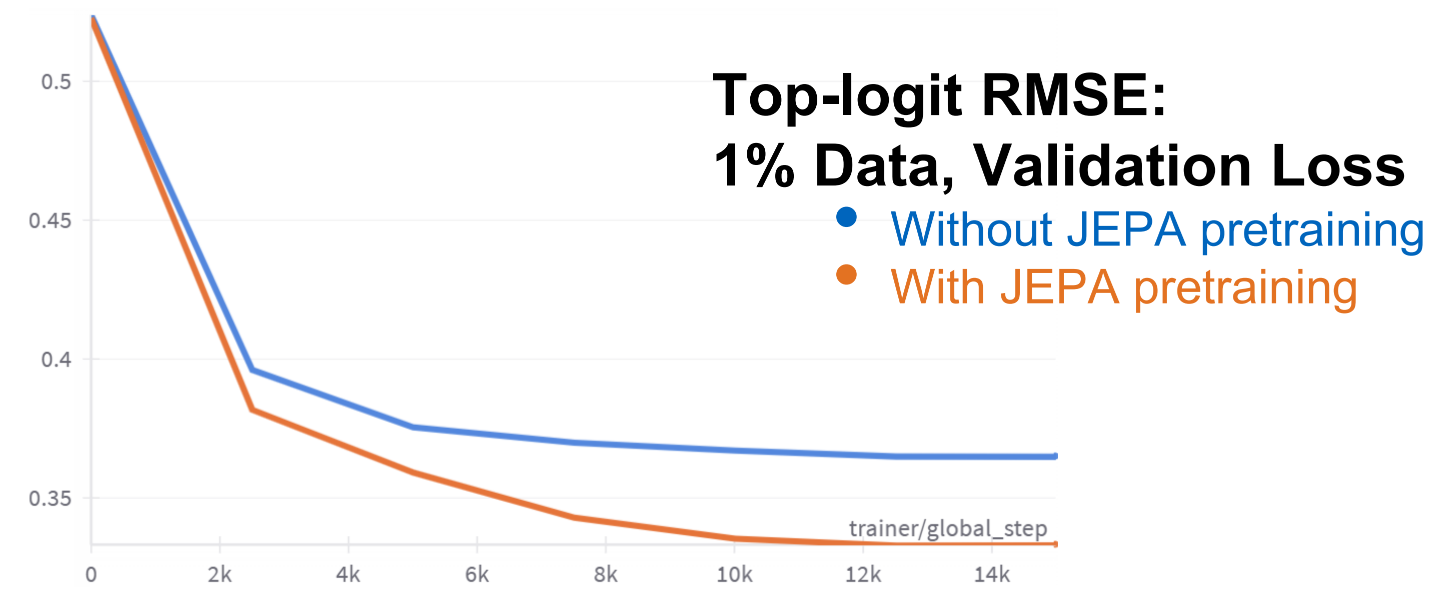
- Diversity is captured by **k=3-5**, with predictions collapsing after.
- LR sweeps showed **finetuning JEPA at 1e-5** to be optimal.
- Transforming coordinate system makes little difference.
- **Learnable logit scale** dynamically balances dual objective.
- **Attention pooling** outperforms mean and max pooling.

Results

- Validation **top-logit RMSE** (rad, lower is better).
- Reflects inference-time performance (since logit is learned).

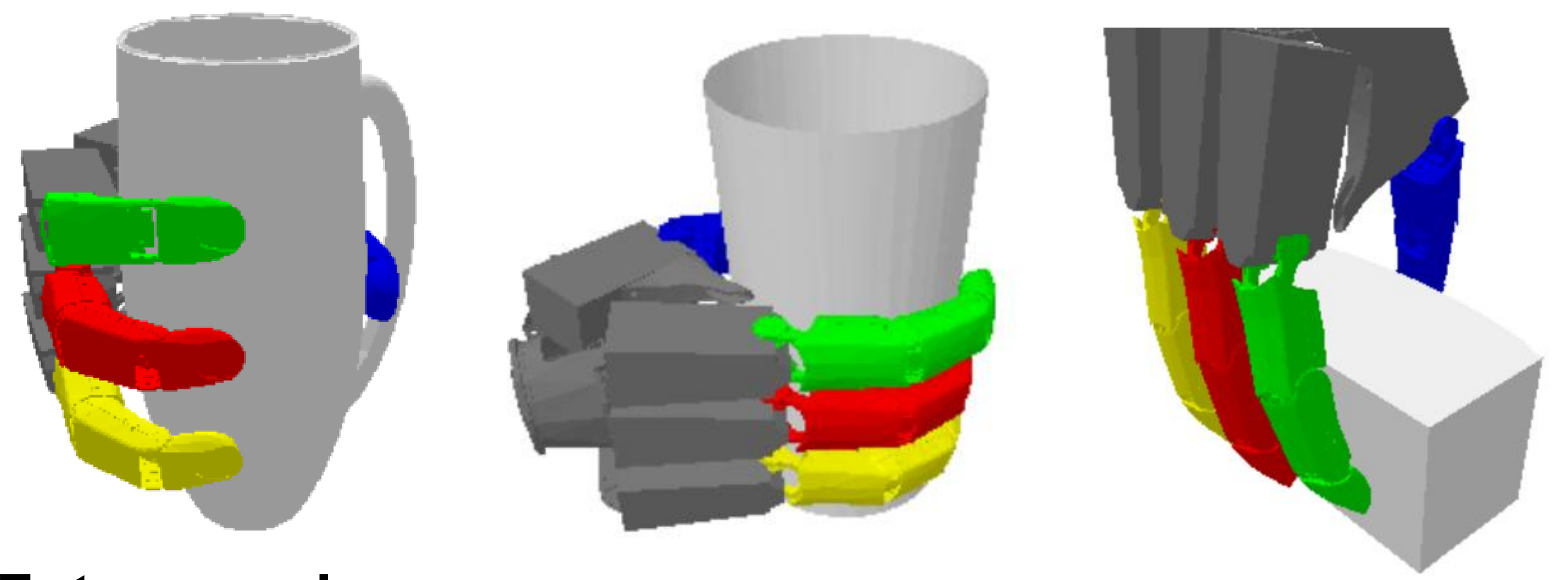
Train split	Scratch	JEPA	Δ (rel.)
1%	0.363 ± 0.002	**0.335 ± 0.003**	+7.7%
10%	0.335 ± 0.003	**0.303 ± 0.009**	+9.6%
25%*	0.331	**0.274**	+17.2%
100%*	**0.232**	0.238	-2.8%

*Single long run; each trained until models plateaued.



Conclusions

- **Tiny data:** JEPA helps, improvements are limited by extreme scarcity.
- **Moderate data**: JEPA yields highest relative gain, "sweet spot" for SSL.
- **Large Data:** No meaningful benefit, scratch fine-tuning suffices.
- **Learning speed:** JEPA pretraining enables quicker convergence in all data regimes.



Future work

- Geometric embeddings, through e.g. DINO
- Ablate different SSL strategies and techniques