Reducing Data in Visual AI: I-JEPA Data Efficiency

Maksim Plotnikov (mplotnikov@tudelft.nl)

CSE3000 Research Project | Supervisors: Jan van Gemert, Alex Manolache, Petter Reijalt

Context & The Problem

Motivation: Self-supervised learning (SSL) is a go-to paradigm for training visual foundation models as it bypasses the need for manual labels. Still, the best models are trained by Big Tech on massive, *proprietary* datasets. This is a blocker for universities and independent labs.

Goal: Measure and improve the low-data efficiency of Self-Supervised Learning (SSL), focusing on I-JEPA.

The SSL Trade-off:

- **Discriminative** (MoCo, DINO): Known to need heavy data augmentation to avoid collapse.
- **Generative** (MAE): Operates in pixel space, over-indexing on trivial, low-level details.
- **I-JEPA:** Predicts in *latent* space, avoiding both pitfalls — an interesting low-data candidate.

Research Questions

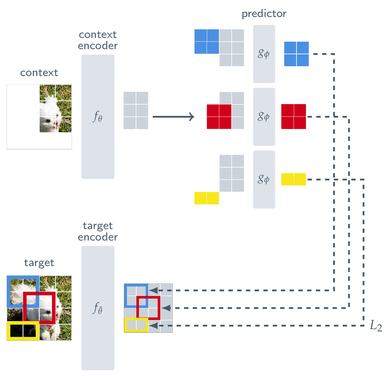
RQ1. How does I-JEPA's downstream representation quality change as SSL pretraining data increases in the low-data regime?

RQ2. How does I-JEPA compare to Barlow Twins, MoCo, DINO, and MAE under a shared protocol?

RQ3. Which simple I-JEPA-specific changes improve transfer in this setting?

I-JEPA Refresher

Motivation: I-JEPA avoids negative pairs and heavy augmentations, making it an ideal candidate for testing whether semantic representations can be learned in smaller data regimes.



Core Idea: Mask an image, encode the visible context, and predict missing regions in *latent space* (not pixels).

SSL Methods Compared

Method	Category	Key Mechanism
Barlow Twins	Contrastive	Redundancy reduction
MoCo, DINO	Discriminative	Matching representations
MAE	Generative	Masked pixel reconstruction
I-JEPA	**Generative (Latent)**	**Representation prediction**

Methodology Pipeline

A **shared downstream evaluation** pipeline with nested, class-stratified dataset sizes makes all findings directly comparable across the models.

1. Stratified Splits — Tiny ImageNet: (1k, 2k, 4k, ... , 100k)

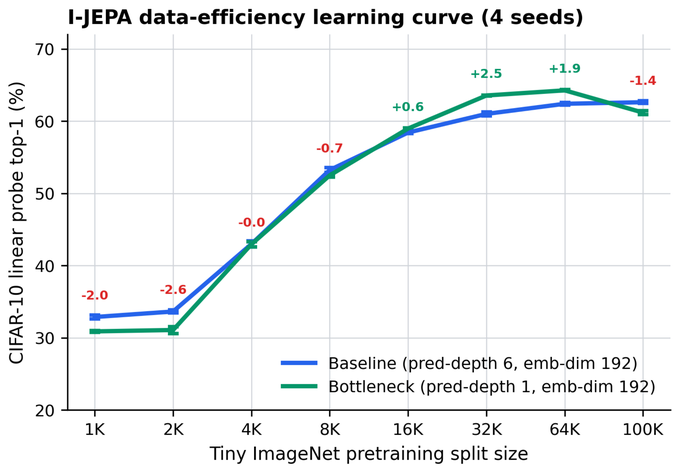
→ **2. Pre-Training** — ViT-Tiny/8, 300 epochs

→ **3. Linear Probe** — CIFAR-10, 100 epochs

→ **4. Evaluation** — CIFAR-10 test set accuracy

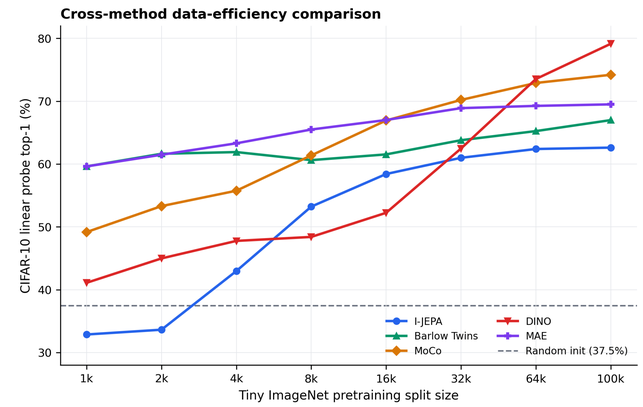
RQ1 — Data-Efficiency Curve

I-JEPA does learn transferable features from restricted data: downstream accuracy rises steeply from ∼**33% (1k)** to ∼**59% (16k)**, slowing and plateauing on larger splits.



RQ2 — Cross-Method Comparison

Under the shared evaluation protocol, all four comparison SSL methods outperform I-JEPA across most of the split range. I-JEPA has a faster early data-efficiency ramp, but starts lower at 1k and plateaus earlier.



RQ3 — Predictor Bottleneck

Among all modifications tested, predictor depth is the most promising direction. Shallow predictor (depth 1 vs. 6) improves transfer at 16k–64k, but harmful or neutral at other splits.

Notably, it alters the training dynamic, with 100k split reaching its transfer peak of ∼65% on epoch 100 (of 300 total), outperforming the baseline. Similar improvements were observed on a bigger backbone ViT-Small.

Depth sweep (32k)

Depth	Top-1
1	**63.6**
2	63.5
3	62.5
6	61.5

Shallower → better

100k trajectory

Epoch	Top-1
50	62.6
100	**65.0**
150	64.3
200	63.2
250	61.7
300	61.2

Peaks at ep100

ViT-Small (△)

Split	BL	BN	△
1k	32.29	29.32	−3.0
2k	32.24	30.45	−2.0
4k	42.83	44.04	+1.2
8k	54.46	52.38	−2.1
16k	58.84	61.52	+2.7
32k	**62.58**	**68.81**	**+6.2**
64k	67.01	70.06	+3.0
100k	69.29	70.28	+1.0

Conclusions

I-JEPA can learn from restricted data, but its performance is lacking compared to other SSL methods tested. Shallow predictor is the clearest direction, but the improvement is data-scale dependent. As such, scheduling and mask tuning experiments are promising next steps.