TU Delft

Reducing Data for Vision Foundation Models

Data-Efficiency of Self-Supervised Learning with DINO Multi-Crop

Leonid Margulis

Supervisors: Jan van Gemert, Alex Manolache, Petter Reijalt · EEMCS, Delft University of Technology

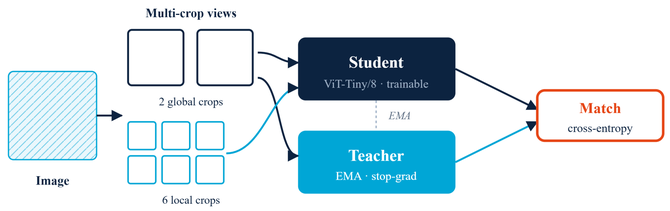
Multi-crop's benefit at sub-ImageNet scale is delayed — not absent — and the optimal N_{local} depends on the downstream task category.

1. Why it matters

- Modern SSL vision models require ImageNet-scale data (≈1.28M images) — out of reach for smaller research groups.
- These are the very groups who **might benefit most** from a cheaper label-free alternative.
- Whether DINO's design choices transfer to **sub-ImageNet pools** (≤100K images, 64×64) is largely untested.
- DINO is the **foundation of DINOv2** — today's standard for visual foundation models.

2. Background & key terms

- **DINO** = Self-Distillation with NO labels (Caron et al., 2021). Student + teacher learn from cropped views of the same image.
- **ViT-Tiny/8** = small Vision Transformer, 5.7M params, 64×64 images as 8×8 patches.
- **Multi-crop** = student sees 2 global + N_{local} local crops; teacher sees only the 2 globals. **The lever we ablate.**
- **Linear probe** = frozen backbone + linear classifier. Measures representation quality without updating the backbone.
- **VTAB-1k** = 19-task transfer benchmark across natural, specialized, and structured categories.



3. Research questions

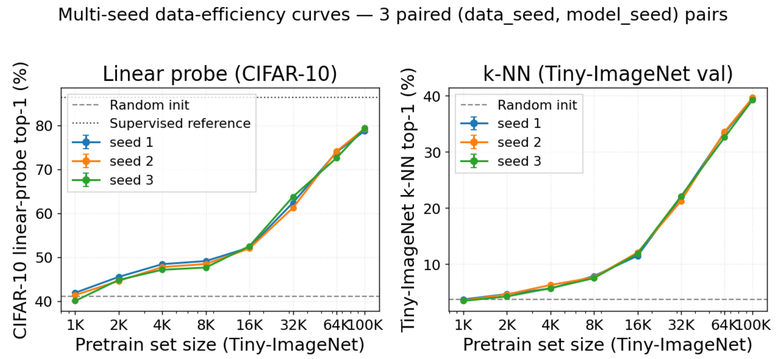
We address three questions:

1. How does DINO's representation quality scale with pretraining-set size between 1K and 100K Tiny-ImageNet images, evaluated on CIFAR-10 linear probe, Tiny-ImageNet k-NN, and VTAB-1k transfer?
2. How does the number of local crops in multi-crop augmentation (N_{local}) interact with pretraining-set size and training duration at this scale?
3. Does the N_{local} choice affect downstream VTAB-1k transfer, and does the effect depend on task category (natural, specialized, structured)?

4. Setup

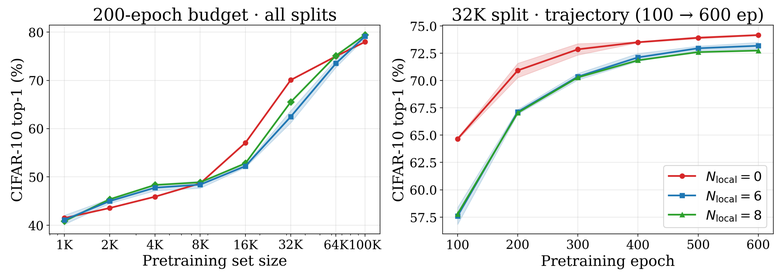
- Pretrain ViT-Tiny/8 with DINO on **8 nested Tiny-ImageNet subsets** (1K → 100K) at 64×64. **Multiseed**: 3 paired seeds for the main data-efficiency curve
- **Three evaluation axes:** CIFAR-10 linear probe · Tiny-ImageNet weighted k-NN · VTAB-1k transfer (19 tasks).
- **200 epochs** (all 8 splits)($N_{\text{local}} \in \{0, 4, 6\}$) and **32K × 600 epochs** ($N_{\text{local}} \in \{0, 6, 8\}$)

5. Q: How does DINO scale from 1K to 100K images? A: Monotonically — but 1K offers no benefit



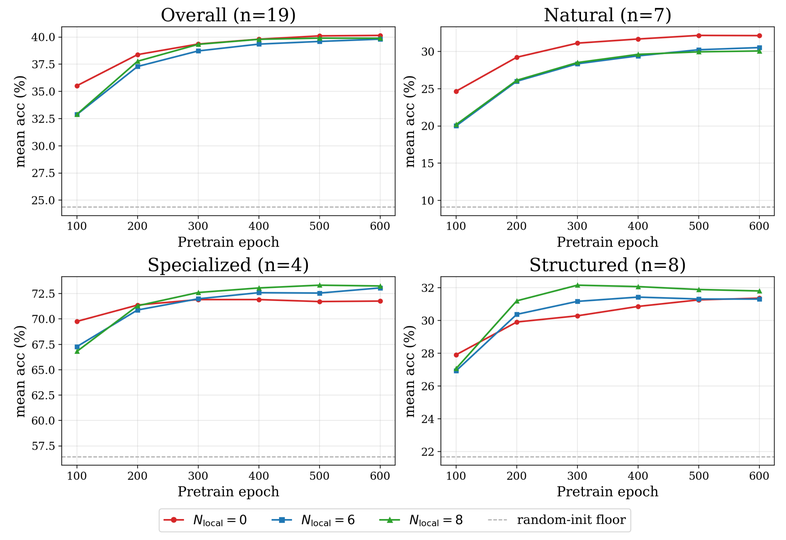
- At 1K, DINO matches random init (41.10 ± 0.93% vs 41.09%).
- From 2K onward, accuracy rises monotonically.
- At 100K: **79.14%** probe (≈92% of the supervised reference 86.31%), 39.47% k-NN, 42.50% VTAB-1k mean.

6. Q: When does multi-crop (N_{local}=6) actually help? A: Not under tight compute — but the lead emerges with extended training



- **200-epoch budget (left):** N_{local}=0 matches or beats N_{local}=6 at medium splits; peak **+7.6 pp** at 32K. U-shape; at 100K the ordering reverses and N_{local}=6 leads.
- **32K × 600 ep (right):** the probe gap narrows from **+3.79 pp** → **+0.98 pp** by ep 600; k-NN ordering even reverses at ep 400 (not shown).
- N_{local}=8 never beats N_{local}=6 — no benefit past six local crops.
- **Multi-crop's benefit at sub-ImageNet scale is delayed, not absent.**

7. Q: Does N_{local} choice affect downstream transfer uniformly? A: No — it depends on task category



- **Natural (7 tasks):** N_{local}=0 wins by +1.6–2.1 pp (svhn ≈ +11 pp).
- **Specialized (4 tasks):** N_{local}=6/8 win by +1.3–1.5 pp (resisc45 ≈ +5 pp).
- **Structured (8 tasks):** all three settings within ≈0.5 pp — effectively tied.

8. Conclusions & limitations

We recommend N_{local}=0 with 200 epochs for compute-constrained natural-image targets, and N_{local}=6 or N_{local}=8 with 600+ epochs for specialized targets. DINO's multi-crop recipe is not one-size-fits-all at sub-ImageNet scale — the right choice emerges from training budget and downstream task category, not the ImageNet default.

Limitations

- Extended training is partly single-seed; gaps within ±2 pp are ordering signals, not confirmed magnitudes.
- Tested at 64×64 with ViT-Tiny/8 only — larger resolutions / backbones untested.
- Downstream scope = linear probe + k-NN + VTAB-1k; no detection or segmentation.