Reducing Data for Vision Foundation Models

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

Leonid Margulis[1]

Supervisors: Jan van Gemert[1] **(Responsible Professor), Alex Manolache**[1]**, Petter Reijalt**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Leonid Margulis
Final project course: CSE3000 Research Project
Thesis committee: Jan van Gemert (Responsible Professor), Alex Manolache, Petter Reijalt

Abstract

Self-supervised learning (SSL) lets computer vision models learn from unlabelled image datasets. Most DINO [1] benchmarks pretrain on ImageNet [2] — a million-image dataset that takes days of multi-GPU training per run, out of reach for the rapid iteration cycles smaller research groups rely on. This leaves practitioners with smaller datasets unsure whether DINO is worth running, or which of its design choices still hold at this scale.

We pretrain a small Vision Transformer (ViT-Tiny/8 [3]) using DINO on Tiny-ImageNet [4] subsets from 1K to 100K images at 64×64 resolution, evaluated on downstream classification tasks. Downstream accuracy grows steadily with pretraining-set size and approaches the accuracy of a fully supervised baseline at the largest scale.

Our main contribution is a multi-crop ablation across data scale, training duration, and downstream task category. We find that multi-crop's benefit at sub-ImageNet scale is delayed rather than absent, and that the optimal multi-crop count depends on the downstream task category — no single setting wins across all tasks.

These findings show that the canonical DINO recipe does not transfer cleanly to sub-ImageNet scale. We recommend choosing the multi-crop count based on training budget and downstream task type, rather than copying the ImageNet default.

1 Introduction

Motivation. Progress in computer vision has long depended on large labelled datasets [2], yet labelled image datasets are expensive — per-image annotation often requires domain experts. Self-supervised learning (SSL) aims to bypass these label costs by training on raw, unlabelled images: the model is trained to produce consistent representations across differently augmented views of the same image. However, modern visual foundation models often need ImageNet-scale unlabelled data and substantial compute, out of reach for smaller research groups — the very groups who would benefit most from a cheaper label-free alternative.

DINO and Multi-Crop. DINO [1] is a self-supervised method that uses two Vision Transformers (ViT) [3], a student and a teacher, where the student learns to match the teacher's output. It is widely used because its learned features work well on new tasks even when the network is frozen and only a simple linear classifier is trained on top [1; 5]. A central design choice in DINO is *multi-crop* augmentation, inherited from SwAV [6]: for each image, the network sees two large global crops and N_{local} smaller local crops. The canonical recipe [1] validated at ImageNet scale uses 6 small local crops alongside 2 global ones.

Research Gap. Published self-supervised learning (SSL) benchmarks — including DINO — are predominantly validated at ImageNet-1K scale: approximately 1.28 million images at 224×224 pixels. Small-scale SSL has been studied for contrastive methods on datasets such as CIFAR and Tiny-ImageNet [5]. Far less is known about how DINO's representation quality changes as the unlabelled pool grows at sub-ImageNet scale (≤100K images, 64×64 pixels), or whether the design choices validated at full scale still hold. DINO could give practitioners in data-scarce domains a cheaper label-free alternative to expanding their labelled set, but only if it works at this scale and the right multi-crop budget for their downstream task can be identified. We focus on DINO specifically because it is the foundation of DINOv2 [7] — the current standard for visual foundation models.

Our Approach. We pretrain a small Vision Transformer with DINO on Tiny-ImageNet [4] at sub-ImageNet scale, evaluate the learned representations on standard downstream benchmarks, and ablate the multi-crop view count (N_{local}) along two axes — data scale and training duration — isolating each effect with a separate sweep. The training-duration sweep is what lets us distinguish a fundamental property of multi-crop from an artefact of training length.

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)?

Contributions. We make three contributions:

- To our knowledge, the first multi-seed ($n = 3$) data-efficiency curve for DINO at sub-ImageNet scale (1K–100K Tiny-ImageNet, 64×64), measured on CIFAR-10 linear probe, Tiny-ImageNet k-NN, and VTAB-1k transfer across 19 tasks in three categories.
- A multi-crop ablation across data scale and training duration. At 200 epochs, $N_{\text{local}} = 0$ matches or beats canonical $N_{\text{local}} = 6$ in the medium-data range, with a CIFAR-10 probe gap of up to 7.6 pp at the 32K split (single-seed $N_{\text{local}} = 0$ vs. three-seed $N_{\text{local}} = 6$ mean). Extending training to 600 epochs at 32K closes the probe gap and partially reverses the k-NN ordering. $N_{\text{local}} = 8$ provides no benefit over $N_{\text{local}} = 6$ on the in-domain probe at this scale (single-seed for $N_{\text{local}} = 8$, two paired seeds for $N_{\text{local}} = 6$ at extended training).
- Downstream evidence that the multi-crop choice depends on task category: $N_{\text{local}} = 0$ wins natural-category VTAB tasks by $+1.6$–2.1 pp; $N_{\text{local}} = 6$ ($n = 2$ paired seeds) and $N_{\text{local}} = 8$ (single seed) both lead specialized tasks by $+1.3$–1.5 pp over $N_{\text{local}} = 0$; on structured tasks all three settings are within ≈ 0.5 pp (single-seed for $N_{\text{local}} = 0$ and $N_{\text{local}} = 8$ in this VTAB protocol; ordering signal, not confirmed magnitude). We recommend choosing N_{local} based on training budget and downstream task type rather than copying the ImageNet default.

2 Related Work

SSL at reduced data scale. Self-supervised learning(SSL) at sub-ImageNet pretraining scales has been studied primarily through end-to-end protocol comparisons. Contrastive learning shows diminishing returns beyond $\sim$500K natural images [5]. Across SSL families compared at 1%, 10%, and 100% of ImageNet, distillation methods such as DINO [1] degrade faster than masked-image-modelling methods on smaller pools [8]. Earlier work on SSL evaluation protocols also showed that probe-protocol choices substantially shift cross-method rankings [9]. Unlike these end-to-end comparisons, we isolate a single DINO [1] design lever — the multi-crop view budget — across data scale and training duration while holding the rest of the pipeline fixed.

Data-efficient Vision Transformers. Vision Transformers originally required large-scale supervised pretraining to reach competitive accuracy [3]. Subsequent work showed that small ViTs can train from scratch on ImageNet-1K labels [10], and that architectural modifications reduce the labelled-data requirement further [11]. These advances all rely on image labels for the target domain; we instead pretrain DINO [1] without labels on a sub-ImageNet pool and evaluate transfer to a separate downstream dataset.

Multi-crop augmentation studies. Multi-crop augmentation feeds additional small local crops to the student while restricting the teacher to global views; it was introduced by SwAV [6] and adopted by DINO [1] with six local crops at 96×96 pixels under ImageNet-scale pretraining. Subsequent foundation-model recipes such as DINOv2 continue to rely on multi-crop augmentation [7]. These works establish multi-crop's value at ImageNet scale and full resolution; we instead ablate the local-crop count in DINO [1] across sub-ImageNet pool sizes, at 64×64 input resolution, and across a training-time trajectory.

3 Methodology

DINO self-distillation. DINO [1] trains an image encoder without labels using *self-distillation*: a learning setup where a student network is trained to mimic the output of a teacher network that shares the same architecture. Both networks receive different augmented views (random crops) of the same image and pass them through a small projection head, producing a probability distribution over $K = 4096$ output dimensions (the DINO default), treated as soft cluster assignments. The student is updated to match the teacher's distribution, which encourages it to produce similar representations for different views of the same image. The teacher is not trained directly — its weights are a slowly-updated exponential moving average (EMA) of the student. To prevent the model from collapsing to a constant output, the teacher's distribution is sharpened with a low softmax temperature and centred by subtracting a running average of teacher outputs across the batch. This consistency-between-views objective is what lets DINO [1] learn transferable features from unlabelled images.

Multi-crop augmentation. Multi-crop [6], the design choice we ablate, generates $2 + N_{\text{local}}$ views per source image (Figure 1). We sample two global crops at 64×64 (area scale $[0.4, 1.0]$) and N_{local} local crops at 32×32 (area scale $[0.05, 0.4]$), upsampling each local crop to 64×64 before it enters the network. The student processes every crop; the teacher processes only the two global crops. The loss pairs every student view (local or global) with each teacher global view, excluding the identical view — so each student crop is matched against both teacher globals, enforcing invariance to scale and position. Varying N_{local} changes how much supervision the student receives from local views; this is the lever our ablations sweep.

Figure 1: Multi-crop augmentation. From each source image we sample two *global* crops at 64×64 and N_{local} *local* crops at 32×32, upsampling each local crop to 64×64 before it enters the ViT. The student network processes every crop; the teacher network processes only the two global crops. The loss pairs each student crop with each teacher global crop, encouraging the student to produce similar representations for views that differ in scale and position. *The figure shows the source image, one representative global crop, and one local crop after upsampling to* 64×64.

Backbone. The encoder is a ViT-Tiny/8 [3; 10] — a small Vision Transformer with about 5.7M parameters. It splits the 64×64 input into $8\times8 = 64$ patch tokens plus a CLS token, processed by self-attention. We use the post-LayerNorm CLS embedding as the image representation. Holding the backbone fixed at ViT-Tiny/8 across all conditions isolates the multi-crop effect from architectural confounds.

4 Experimental Setup

We specify the DINO [1] pretraining protocol, datasets, three evaluation axes, and baselines.

4.1 Pretraining data

We pretrain on Tiny-ImageNet [4]: 100,000 images, 200 classes, native 64×64. We construct 8 nested subsets (1K, 2K, 4K, 8K, 16K, 32K, 64K, 100K), class-stratified by interleaving one image per class per round from a pre-shuffled index list (seed 42). Any subset of size $N \geq 200$ then contains all 200 classes with at most one image of imbalance. Figure 2 shows samples. The two datasets come from disjoint sources, so the probe measures transfer of the DINO [1] representation rather than recall.

4.2 Pretraining protocol

We pretrain every split for the same 200 epochs so dataset size is the only variable. All runs use AdamW [12] at batch size 256 with gradient clipping at norm 3.0; Table 1 lists the

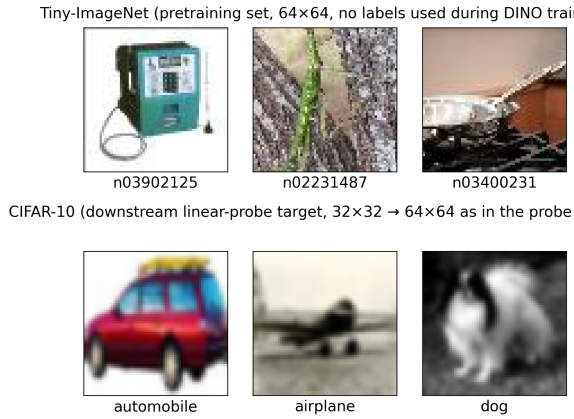

Figure 2: Sample images from the pretraining and downstream datasets. Top: Tiny-ImageNet (pretraining, 64×64, no labels used during DINO training). Bottom: CIFAR-10 (downstream linear-probe target, native 32×32 shown resized to 64×64 as in the probe pipeline). The two datasets come from disjoint sources, so the CIFAR-10 probe measures transfer of the learned representation rather than recall of seen images.

full configuration. Optimiser, learning-rate schedule, weight-decay ramp, EMA momentum, and temperature schedules follow the reference DINO [1] recipe. This protocol drives the main data-efficiency curve and the ablation of the local-crop count $N_{\text{local}} \in \{0, 4, 6\}$ across all eight splits.

Extended-training sub-protocol. To probe how N_{local} interacts with training time, we also run a 600-epoch trajectory on the 32K split with $N_{\text{local}} \in \{0, 6, 8\}$. We stretch the same recipe to 600 epochs, rescaling the warmup, cosine LR, weight-decay, EMA, and teacher-temperature ramps. This gives 3× more updates per image at fixed data scale. We use three paired seeds for $N_{\text{local}} = 0$, two paired seeds for $N_{\text{local}} = 6$, and a single seed for $N_{\text{local}} = 8$.

4.3 Evaluation axes

CIFAR-10 linear probe. We follow the standard linear-probe protocol for self-supervised representations [9]: freeze the backbone, train a fresh $\text{Linear}(192, 10)$ head with stochastic gradient descent (SGD; momentum 0.9, no weight decay) and a cosine learning rate (LR) schedule over 100 epochs. The base LR is scaled with the linear-rule [13; 14]. Transforms are Resize(64) with ImageNet normalisation. We omit the train-side RandomHorizontalFlip of the reference DINO [1] probe to enable feature caching. The metric is best top-1 test accuracy.

Tiny-ImageNet weighted k-NN. The weighted k-NN protocol of DINO [1] trains no parameters. We extract post-LayerNorm CLS features for all Tiny-ImageNet train and validation images and L2-normalise them. We then score class c for validation feature q using its top-k neighbours $\mathcal{N}_k(q)$ as

$$s_c(q) = \sum_{i \in \mathcal{N}_k(q)} \exp\left(\frac{q^\top f_i}{\tau}\right) \mathbf{1}[y_i = c], \tag{1}$$

where f_i is the L2-normalised feature of training image i, y_i its class label, $q^\top f_i$ the dot-product (cosine) similarity, τ a temperature that sharpens the neighbour weights, and $\mathbf{1}[\cdot]$ the indicator function (1 if the bracketed condition holds, 0 otherwise). We use $k = 20$ and $\tau = 0.07$ following DINO [1], evaluated on the 10,000-image Tiny-ImageNet validation split. We log every 20 epochs.

Table 1: Hyperparameter summary for the uniform 200-epoch DINO pretraining protocol and the three evaluation axes. Rows marked ‡ apply to the 32K × 600-epoch extended-training sub-protocol.

Component	Value
Backbone	ViT-Tiny/8 (64×64)
Projection head dims	192 → 2048 → 2048 → 256 → 4096
Pretrain batch	256
Pretrain epochs (main)	200 (uniform across all 8 splits)
Pretrain epochs (extended)‡	600 (32K split only)
Pretrain LR	5×10^{-4} / 10-ep warmup / cos to 10^{-6}
Weight decay (start / end)	0.04 / 0.4 (cosine)
EMA momentum (start / end)	0.996 / 1.0 (cosine)
Teacher / student temperature	0.04→0.07 / 0.1
Teacher temp warmup	30 epochs
Multi-crop	2×64 + N_{local}×32→64
N_{local} ablation (200 ep)	$\{0, 4, 6\}$
N_{local} ablation (extended, 600 ep)‡	$\{0, 6, 8\}$
CIFAR-10 probe / VTAB-1k probe	SGD m=0.9, no weight decay, cosine LR
CIFAR-10 probe epochs	100
VTAB-1k probe epochs	90
k-NN parameters	$k = 20$, $\tau = 0.07$

VTAB-1k transfer. VTAB-1k [15] tests whether the CIFAR-10 trend generalises across natural, specialized, and structured categories. We use the canonical `train800val200` split (1000 training images per task) and evaluate on each task's held-out test set across all 19 tasks. We train per-task linear classifiers with SGD (momentum 0.9, no weight decay), cosine LR, batch size 256, 90 epochs, and aggregate per category and overall.

4.4 Baselines

As a random-init baseline, we evaluate a ViT-Tiny/8 with weights left at their default random initialisation — without pretraining — under the same probe, k-NN, and VTAB-1k linear-probe protocols. This isolates the contribution of DINO [1] pretraining from the architecture and probe pipeline.

For a supervised reference of 86.31%, we train the same ViT-Tiny/8 end-to-end on CIFAR-10 for 200 epochs with AdamW. Optimisation uses peak LR 5×10^{-4}, weight decay 0.05, batch size 256, 10-epoch warmup, cosine decay to 10^{-6}, and label smoothing 0.1. Augmentation uses RandomCrop(64, padding 4), RandomHorizontalFlip p=0.5, ColorJitter strength 0.4, and Bicubic Resize, with CIFAR-10 dataset normalisation statistics. This contextualises how

much of the supervised accuracy (86.31%) frozen DINO [1] features recover with only a linear head.

4.5 Deviations from the reference DINO recipe

We list our compromises against the reference DINO [1] recipe:

- Input resolution 64×64; backbone ViT-Tiny; projection-head output 4096 (paper: 224×224, ViT-S/B, 65,536).
- Local crops upsampled to 64×64 rather than positional-embedding interpolation.
- Uniform 200-epoch budget across splits, replacing the variable per-split schedule of [8].
- $N_{\text{local}} = 8$ not run at the uniform 200-epoch protocol; included only at the extended-training sub-protocol.
- Linear probe trained without train-side horizontal flip for 100 epochs to enable feature caching.
- Single seed for $N_{\text{local}} \in \{0, 4\}$ at uniform 200 epochs and for $N_{\text{local}} = 8$ at the extended-training sub-protocol; two paired seeds for $N_{\text{local}} = 6$ at extended training; three paired seeds for the $N_{\text{local}} = 6$ data-efficiency curve and for the extended-training $N_{\text{local}} = 0$ run.

Our compute budget forced these deviations. We expect the qualitative trends to hold even though absolute numbers may differ from canonical DINO [1] at ImageNet scale.

5 Results

Section 5.1 explains how DINO [1] representation quality scales with pretraining-set size. Section 5.2 shows the scaling transfers downstream. Section 5.3 shows how the number of local crops N_{local} interacts with split size at 200 epochs. Section 5.4 extends training to 600 epochs at the critical 32K split, and Section 5.5 asks whether N_{local} choice matters per downstream task category.

5.1 Experiment 1: Main data-efficiency curve

Question. How does DINO's representation quality scale with pretraining-set size from 1K to 100K Tiny-ImageNet images?

Figure 3 shows CIFAR-10 linear-probe and Tiny-ImageNet weighted k-NN accuracy for DINO [1] as a function of pretraining-set size, measured across three paired random seeds. A random-initialised ViT-Tiny/8 (no pretraining) scores 41.09% probe / 3.76% k-NN, providing a floor for comparison.

At 1K, DINO [1] reaches $41.10 \pm 0.93\%$ probe and $3.57 \pm 0.13\%$ k-NN — within 0.01 pp of random init on probe and 0.19 pp *below* it on k-NN. Pretraining on 1K images therefore provides no measurable benefit at 200 epochs, to our knowledge not previously reported at this scale.

From 2K, both curves rise monotonically. By 100K, DINO [1] reaches $79.14 \pm 0.35\%$ probe and $39.47 \pm 0.19\%$ k-NN — gains of +38.04 and +35.90 pp over 1K, and approximately 92% of the supervised reference of 86.31%. Inter-seed std widens at medium splits (up to 1.27 pp at 32K) but stays small relative to the cross-N_{local} gap (up to 7.60 pp), so seed effects are not the primary variance source.

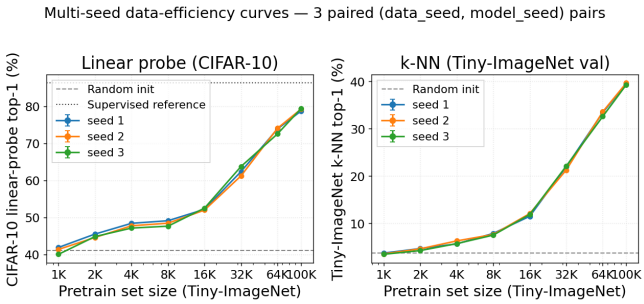

Figure 3: **DINO data-efficiency curve (uniform 200-epoch budget, $N_{\text{local}} = 6$, $n = 3$ paired seeds).** *X-axis:* Tiny-ImageNet pretraining-set size (log scale, 1K–100K). *Y-axis:* Top-1 accuracy (%). *Solid curves with shading:* mean ± std across three paired seeds for CIFAR-10 linear probe (blue) and Tiny-ImageNet weighted k-NN (orange). *Dashed line:* random-initialisation floor (41.09% probe, 3.76% k-NN). *Dotted line:* supervised reference at 86.31%. **Key finding:** at 1K, DINO does not beat random initialisation under this budget; from 2K onward both curves rise monotonically to $79.14 \pm 0.35\%$ probe and $39.47 \pm 0.19\%$ k-NN at 100K.

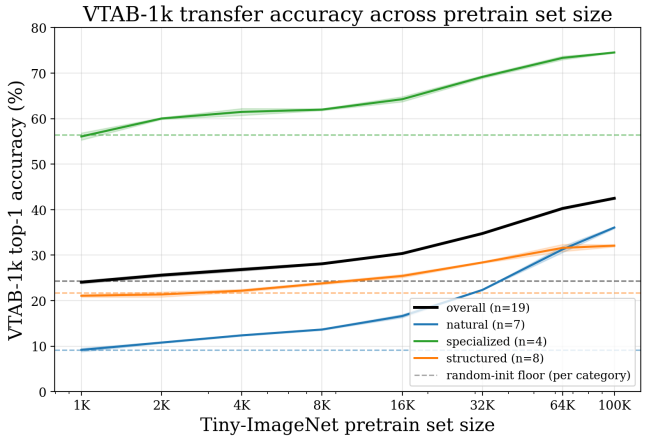

Figure 4: **VTAB-1k downstream transfer vs. pretraining-set size (DINO, $N_{\text{local}} = 6$, 200 ep, $n = 3$ paired seeds).** *X-axis:* Tiny-ImageNet pretraining-set size (log scale). *Y-axis:* Mean top-1 accuracy (%). *Curves:* overall VTAB mean (black), natural (green), specialized (red), structured (blue); shading = ±std across three seeds. *Dashed horizontal lines:* per-category random-init floor (overall 24.37%, natural 9.12%, specialized 56.43%, structured 21.69%). **Key finding:** all three categories and the overall mean rise monotonically from 1K to 100K; the overall mean gains +18.13 pp over the random-init floor and +18.43 pp over the 1K split. DINO@1K matches the random-init floor within ∼ 1 pp in every category, showing that 1K pretraining provides no measurable downstream benefit. `diabetic_retinopathy` (and `patch_camelyon`) saturate near majority-class rates including at random init and are disclosed as artefacts of the linear-probe protocol on imbalanced or binary tasks.

5.2 Experiment 2: VTAB-1k downstream transfer

Question. Does the scaling trend hold on downstream VTAB-1k transfer across natural, specialized, and structured tasks?

Figure 4 shows VTAB-1k [15] mean transfer accuracy

across 19 tasks for the same checkpoints as Section 5.1. The overall mean rises monotonically from 24.07% at 1K to 42.50% at 100K (+18.43 pp). Inter-seed std at the overall-mean level is 0.08–0.34 pp across splits, so the signal is not noise-dominated.

All three categories are monotone from 1K to 100K: natural-image (7 tasks: `caltech101`, `cifar`, `dtd`, `oxford_flowers102`, `oxford_iiit_pet`, `sun397`, `svhn`) rises 9.18% → 36.07% (+26.89 pp); specialized (4 tasks: `eurosat`, `patch_camelyon`, `resisc45`, `diabetic_retinopathy`) rises 56.09% → 74.55% (+18.46 pp); structured (8 tasks) rises 21.08% → 32.09% (+11.01 pp). Compared to the random-init floor (Section 5.2), the gain at 100K is +26.95 pp natural (from 9.12%), +18.12 pp specialized (from 56.43%), and +10.40 pp structured (from 21.69%).

A random-init ViT-Tiny/8 backbone (single seed) scores 24.37% overall mean on VTAB-1k (9.12% natural, 56.43% specialized, 21.69% structured). DINO@1K matches this floor within ∼ 1 pp in every category, mirroring the in-domain result of Section 5.1: pretraining on 1K images provides no measurable downstream benefit over no pretraining. From 2K onward the curves separate from the floor, reaching 42.50% overall at 100K — a +18.13 pp gain over the random-init floor that is consistent with the scaling signal being driven by pretraining rather than the linear-probe pipeline.

Two tasks require disclosure: `diabetic_retinopathy` scores 73.59 ± 0.00% at every split and seed, including random init — the linear probe predicts the majority class regardless of pretraining volume, indicating the frozen ViT-Tiny CLS token carries no discriminative signal for retinal fundus images at this scale. Likewise `patch_camelyon` scores 72.09% at random init on this 2-class task, suggesting near-majority prediction. We retain both in the VTAB mean for consistency but they inflate the specialized-category floor.

5.3 Experiment 3: N_{local} ablation across split sizes at 200 epochs

Question. How does the number of local crops (N_{local}) interact with pretraining-set size under a uniform 200-epoch budget?

DINO [1] multi-crop generates two global crops (full resolution) and N_{local} local crops (32×32 upsampled to 64×64). We compare $N_{\text{local}} \in \{0, 4, 6\}$ across all eight splits under the 200-epoch budget. $N_{\text{local}} = 6$ uses three paired seeds (from Section 5.1); $N_{\text{local}} = 0$ and $N_{\text{local}} = 4$ are single-seed, limiting the strength of per-point comparisons.

Figure 5 shows the curves. The $N_{\text{local}} = 0$ vs. $N_{\text{local}} = 6$ gap follows a U-shape across data volume, peaking at 32K. At 16K, $N_{\text{local}} = 0$ reaches 57.07% probe against 52.24±0.27% for $N_{\text{local}} = 6$ (+4.83 pp). At 32K, the gap is +7.60 pp on probe (70.07% vs. 62.47 ± 1.27%) and +4.82 pp on k-NN (26.59% vs. 21.77 ± 0.44%). Both gaps are several times the inter-seed std of $N_{\text{local}} = 6$, so the ordering is unlikely to be a seed artefact, though formal significance is not established at single seed for $N_{\text{local}} = 0$.

The gap narrows toward the data extremes: at 64K the $N_{\text{local}} = 0$ probe lead is +1.47 pp; at 100K it reverses to −1.16 pp (77.98% vs. 79.14 ± 0.35%), aligning with the multi-crop benefit reported for DINO [1] at ImageNet scale. This 200-epoch snapshot shows that removing local crops matches or leads multi-crop in the medium-data regime (8K–32K). Section 5.4 tests whether the ordering persists under extended training.

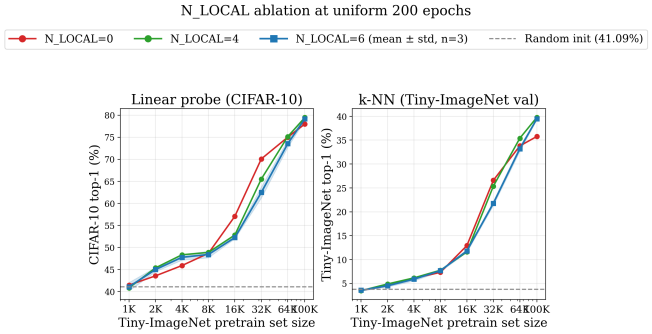

Figure 5: **N_{local} ablation across split sizes at 200 epochs (DINO, uniform 200-epoch budget).** *X-axis:* Tiny-ImageNet pretraining-set size (log scale). *Y-axis:* Top-1 accuracy (%). *Curves:* $N_{\text{local}} = 0$ (no local crops, single seed, solid), $N_{\text{local}} = 4$ (single seed, dashed), $N_{\text{local}} = 6$ (three paired seeds, mean ± std shading). Left panel: CIFAR-10 linear probe. Right panel: Tiny-ImageNet k-NN. **Key finding:** $N_{\text{local}} = 0$ leads $N_{\text{local}} = 6$ by up to +7.60 pp probe at 32K under the 200-epoch budget, forming a U-shaped gap that reverses at 100K. $N_{\text{local}} = 0$ and $N_{\text{local}} = 4$ are single-seed; gaps relative to the multi-seed $N_{\text{local}} = 6$ are informative but lack per-seed confirmation for those two values.

5.4 Experiment 4: In-domain trajectory at 32K × 600 epochs

Question. Does the $N_{\text{local}} = 0$ lead at 32K reflect an inherent property of multi-crop at this scale, or does extended training close the gap? We pretrain $N_{\text{local}} \in \{0, 6, 8\}$ on the 32K split for 600 epochs and evaluate at six milestones (epochs 100, 200, 300, 400, 500, 600). $N_{\text{local}} = 0$ uses three paired seeds, $N_{\text{local}} = 6$ uses two paired seeds, and $N_{\text{local}} = 8$ uses a single seed due to compute constraints.

Figure 6 shows the evolution. At epoch 200, $N_{\text{local}} = 0$ reaches 70.93 ± 0.55% probe and 28.05 ± 0.42% k-NN, against 67.14 ± 0.21% probe and 27.00 ± 0.40% k-NN for $N_{\text{local}} = 6$ (+3.79 pp probe, +1.06 pp k-NN). Both curves continue improving. By epoch 600, $N_{\text{local}} = 0$ reaches 74.17 ± 0.04% probe and 32.18 ± 0.30% k-NN; $N_{\text{local}} = 6$ reaches 73.19 ± 0.31% probe and 32.92 ± 0.08% k-NN. The probe gap narrows to +0.98 pp; on k-NN it reverses — $N_{\text{local}} = 6$ leads by +0.74 pp. The k-NN reversal first appears at epoch 400 (−0.35 pp) and grows monotonically thereafter. $N_{\text{local}} = 8$ (single seed) lies below $N_{\text{local}} = 6$ at all six probe milestones (e.g., 72.75% vs. 73.19±0.31% at epoch 600) and is below on k-NN; eight local crops provide no benefit over six at this scale.

This evolution reframes the 200-epoch result: the $N_{\text{local}} = 0$ lead is a snapshot in a closing gap, not a stable ordering. Multi-crop's benefit in DINO [1] is delayed at sub-ImageNet scale rather than absent. The metric asymmetry — k-NN reverses, probe still favours $N_{\text{local}} = 0$ by 0.98 pp — suggests multi-crop primarily improves local feature discrimina-

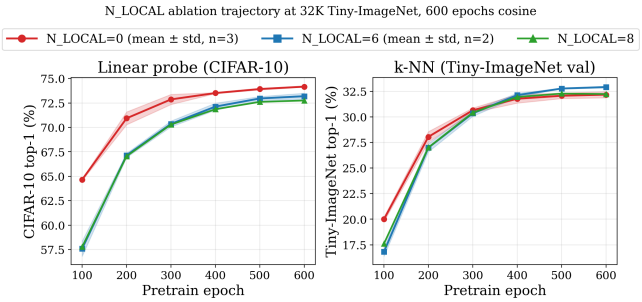

Figure 6: **Training trajectory at 32K Tiny-ImageNet, epochs 100–600 (DINO, $N_{\text{local}} \in \{0, 6, 8\}$).** *X-axis:* Pretraining epoch. *Y-axis:* Top-1 accuracy (%). $N_{local} = 0$*:* multi-seed mean ± std ($n = 3$ paired seeds, shaded band). $N_{local} = 6$*:* multi-seed mean ± std ($n = 2$ paired seeds, shaded band). $N_{local} = 8$*:* single seed (dashed line). Left panel: CIFAR-10 linear probe. Right panel: Tiny-ImageNet k-NN. **Key finding:** the $N_{\text{local}} = 0$ probe lead of $+3.79$ pp at epoch 200 narrows to $+0.98$ pp at epoch 600, and the k-NN ordering reverses at epoch 400 ($N_{\text{local}} = 6$ leads by $+0.74$ pp at epoch 600). Multi-crop benefit is delayed, not absent. $N_{\text{local}} = 8$ lies below $N_{\text{local}} = 6$ throughout; $N_{\text{local}} = 6$ has two paired seeds at this protocol, $N_{\text{local}} = 8$ a single seed.

tion (k-NN) over globally linearly-separable representations (probe). Since $N_{\text{local}} = 8$ is single-seed here, the comparison against $N_{\text{local}} = 8$ is an ordering signal; the $+0.74$ pp k-NN reversal for $N_{\text{local}} = 6$ is now observed across two paired seeds.

5.5 Experiment 5: Downstream transfer by task category at 32K × 600 epochs

Question. Does N_{local} choice affect downstream transfer uniformly across task categories? We evaluate VTAB-1k on the 32K × 600-epoch checkpoints ($N_{\text{local}} \in \{0, 6, 8\}$) at the six milestones. $N_{\text{local}} = 6$ uses two paired seeds (d42_m0, d43_m1); $N_{\text{local}} = 0$ and $N_{\text{local}} = 8$ are single-seed.

Figure 7 shows per-category accuracy at epoch 600. The overall VTAB mean converges to $\approx$ 40% for all three configurations, but the category breakdown reveals a task-dependent split. For **natural-image tasks** (7 tasks), $N_{\text{local}} = 0$ leads (32.1% vs. 30.5% for $N_{\text{local}} = 6$ ($n = 2$ mean) and 30.1% for $N_{\text{local}} = 8$, a margin of +1.6–2.1 pp). The clearest per-task advantage is `svhn` (+11 pp over $N_{\text{local}} = 6$), with all other natural tasks within ±1 pp. For **specialized tasks** (4 tasks), $N_{\text{local}} = 6$ and $N_{\text{local}} = 8$ lead (73.1% and 73.3% vs. 71.8% for $N_{\text{local}} = 0$, a margin of +1.3–1.5 pp). `resisc45` shows the clearest gap ($N_{\text{local}} = 8$ ahead of $N_{\text{local}} = 0$ by ≈ 5 pp). For **structured tasks** (8 tasks), the three settings are roughly tied within ≈ 0.5 pp at the category mean ($N_{\text{local}} = 8$: 31.8%, $N_{\text{local}} = 0$: 31.4%, $N_{\text{local}} = 6$ ($n = 2$): 31.3%). Per-task, the largest structured-task gap is `clevr_dist`, where $N_{\text{local}} = 0$ leads $N_{\text{local}} = 6$ by +6.3 pp; `dsprites_loc`, `dsprites_ori`, and `smallnorb_ele` favour multi-crop (`dsprites_loc`: $N_{\text{local}} = 8$ ahead of $N_{\text{local}} = 0$ by ≈ 4 pp). Two tasks saturate regardless of N_{local}: `diabetic_retinopathy` at 73.59% (Section 5.2) and `patch_camelyon` near 78%. Category means track the in-domain trajectory across epochs 100–600, with gaps remaining small (≤ 2 pp).

The pattern indicates N_{local} choice is not universal at this scale: removing local crops benefits natural-image transfer; restoring them benefits specialized tasks; structured tasks are essentially neutral. With overall VTAB means close to equal ($\approx 40\%$), N_{local} should be matched to the target downstream category. Since $N_{\text{local}} = 0$ and $N_{\text{local}} = 8$ are single-seed (and $N_{\text{local}} = 6$ is $n = 2$), the per-category margins ($\approx$ 1–2 pp) are indicative rather than statistically confirmed.

6 Discussion

We measured DINO [1] across three axes at sub-ImageNet scale: pretraining-set size, multi-crop view budget (N_{local}), and training duration. The following subsections interpret each axis in turn — how representation quality scales with data, why multi-crop's benefit is delayed at this scale, how downstream gains differ by task category, and where N_{local} saturates — and close with practical guidance and a refinement of our initial hypothesis.

6.1 The data-efficiency curve

At 1K under 200 epochs, DINO [1] matches random initialisation on probe and falls slightly below it on k-NN: the consistency loss cannot escape the random-init basin from two global crops on so few distinct images. From 2K the curve rises monotonically and is still climbing at 100K, so practitioners can expect further gains beyond this cap. At 100K, frozen DINO features recover about 92% of the supervised CIFAR-10 reference with no labels — evidence that sub-ImageNet self-supervision captures most of the supervised signal on CIFAR-10 at this resolution.

6.2 What "delayed" multi-crop means

Section 5.4 shows multi-crop benefit arrives later than at ImageNet scale. A token-budget argument may explain why. At 64×64 input with patch size 8, a 32×32 local crop covers a 4×4 grid of source patches; after upsampling to 64×64 the ViT processes 64 tokens that encode information from only 16 unique source patches. At short budgets the student receives enough supervision from two global crops, and the local-global consistency loss is redundant with global-global on these information-sparse views. At longer budgets the model has extracted most of the signal from global pairs, and local crops become a complementary supervision source. The same delayed-utility pattern matches multi-crop's role at ImageNet [1]: 1.28M images at 224×224 for 300+ epochs places each image in a long-budget regime, so local crops help from the start. At 32K Tiny-ImageNet images at 64×64, the model has processed $\sim 40\times$ fewer total sample views by epoch 200 than at ImageNet scale, and multi-crop's benefit only appears once training is extended. Multi-crop, introduced in SwAV [6] and adopted in DINO at ImageNet scale, shifts its threshold of utility when both resolution and data volume are reduced.

An alternative explanation for the closing gap is that $N_{\text{local}} = 6$ accumulates roughly $4\times$ more student forward passes per epoch than $N_{\text{local}} = 0$, so by epoch 600 it has used $\sim 12\times$ more student compute, and the catch-up could reflect

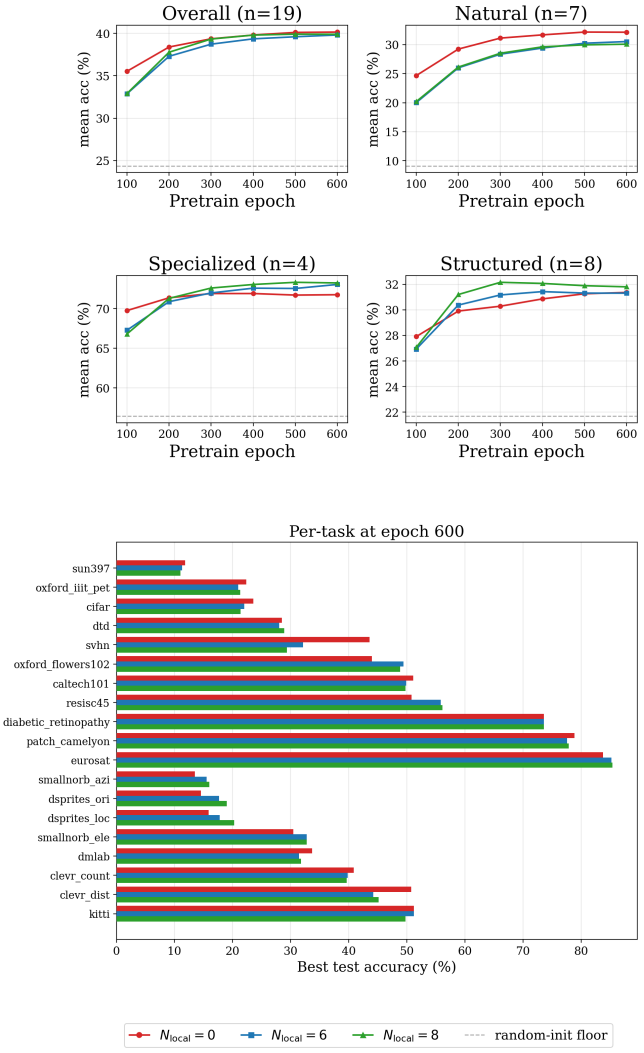

Figure 7: **VTAB-1k transfer by category and N_{local} at 32K Tiny-ImageNet × 600 epochs (DINO).** *Top panel:* Per-category mean accuracy vs. epoch (100–600) for $N_{\text{local}} \in \{0, 6, 8\}$; grey dashed lines mark the per-category random-init floor from Section 5.2. *Bottom panel:* Per-task bar chart at epoch 600, sorted by task; colour-coded by N_{local} value. $N_{local} = 6$*:* mean across two paired seeds (d42_m0, d43_m1). $N_{local} = 0$, $N_{local} = 8$*:* single seed each (d42_m0). **Key finding:** the overall VTAB mean ($\approx$ 40%) is similar for all three N_{local} values at epoch 600, but natural-image tasks favour $N_{\text{local}} = 0$ (+1.6–2.1 pp; `svhn` +11 pp) while specialized tasks favour $N_{\text{local}} = 6/8$ (+1.3–1.5 pp; `resisc45` $\approx$ +5 pp); structured tasks are essentially tied across all three settings ($\approx$ 0.5 pp spread). N_{local} choice should be matched to the intended downstream task category.

compute rather than algorithmic properties of multi-crop. Two observations make pure-compute an unlikely sole driver. First, at 200 epochs $N_{\text{local}} = 6$ already has 4× more student compute yet loses by +3.79 pp on probe and +1.06 pp on k-NN, so "more compute helps" alone does not predict the early gap. Second, the catch-up on the in-domain probe is gradual (+3.79 → +0.98 pp over 400 epochs), not the sharp transition expected if a compute threshold triggered multi-crop's benefit. A compute-matched protocol (e.g., $N_{\text{local}} = 6$ at $\sim 1/4$ the epochs of $N_{\text{local}} = 0$) would isolate the effect; we leave this to future work.

6.3 Task-category interaction

N_{local} choice is not neutral across task types (Section 5.5): $N_{\text{local}} = 0$ leads on natural-image transfer, $N_{\text{local}} = 6/8$ lead on specialized tasks, and structured tasks are tied within ≈ 0.5 pp. The pattern follows how information is distributed in each task. Satellite scenes (`resisc45`, `eurosat`) and abstract patterns (`dsprites_loc`) have no dominant foreground object, so local crops help by encoding spread-out structure. Natural-image tasks like `svhn` and `caltech101` are dominated by a single foreground object, where extra local views do not help — and sometimes hurt — the global representation. This matches DINO's [1] strength at predominantly object-centric ImageNet, but the benefit does not transfer cleanly to specialized or structured benchmarks. The climb from the random-init floor (Section 5.2) is also uneven across categories: natural tasks gain +26.95 pp at 100K, specialized +18.12 pp, and structured only +10.40 pp.

6.4 Saturating N_{local}

Adding local crops beyond $N_{\text{local}} = 6$ provides no measurable benefit at this scale. On the in-domain evaluation of Section 5.4, $N_{\text{local}} = 8$ (single seed) lies below $N_{\text{local}} = 6$ (two paired seeds) on probe at all six milestones (e.g., 72.75% vs. 73.19 ± 0.31% at epoch 600), and is also below on k-NN (32.24% vs. 32.92 ± 0.08%). On downstream VTAB-1k transfer at epoch 600, $N_{\text{local}} = 8$ matches $N_{\text{local}} = 6$ ($n = 2$ mean) to within 0.2 pp on the specialized category mean and exceeds it by ~ 0.5 pp on structured tasks — a modest gap given the single seed for $N_{\text{local}} = 8$. We attribute this to the low token count per local crop: with 16 unique source patches per view, additional crops beyond $N_{\text{local}} = 6$ overlap heavily in information content and the marginal diversity per added view is low. DINO [1] noted at ImageNet scale that gains diminish beyond $N_{\text{local}} = 6$; our results extend this to the sub-ImageNet, low-resolution regime, where the plateau is already at $N_{\text{local}} = 6$.

6.5 Practical guidance

Two practical recommendations follow for DINO [1] at sub-ImageNet scale. First, under a tight compute budget ($\leq$ 200 epochs) with natural-image downstream tasks and medium splits (8K–32K), $N_{\text{local}} = 0$ matches or exceeds the canonical $N_{\text{local}} = 6$ recipe while using 4× fewer student forward passes per batch; at the 100K split this ordering reverses and $N_{\text{local}} = 6$ leads. Second, under a generous budget ($\geq$ 600 epochs) targeting specialized tasks, $N_{\text{local}} = 6$ yields a ~ 1.3 pp category-mean transfer advantage; on structured tasks all three N_{local} settings are within ~ 0.5 pp. The standard $N_{\text{local}} = 6$ recipe is therefore not universally optimal at sub-ImageNet scale — the choice depends on training budget, split size, and downstream task category.

6.6 Hypothesis refinement

Our initial hypothesis was that multi-crop would hurt at small pretraining scale. The 200-epoch ablation appeared to sup-

port this ($N_{\text{local}} = 0$ led by up to $+7.60$ pp probe at 32K). The 600-epoch trajectory refined the conclusion to "delayed, not absent" in-domain, and the VTAB-1k results added a task-category dimension absent from the original hypothesis. A single training-time snapshot can mislead; multi-axis ablations spanning data volume, training budget, and downstream task type are needed to characterise DINO [1] behaviour at sub-ImageNet scale.

7 Limitations

- **Multi-seed coverage.** The main data-efficiency curve uses three paired seeds. At 200 epochs, $N_{\text{local}} = 0$ and $N_{\text{local}} = 4$ are single-seed, so the $+7.60$ pp probe gap at 32K compares a single-seed $N_{\text{local}} = 0$ against the three-seed $N_{\text{local}} = 6$ mean. The 32K $\times$ 600-epoch trajectory uses three paired seeds for $N_{\text{local}} = 0$, two for $N_{\text{local}} = 6$, and one for $N_{\text{local}} = 8$ due to compute constraints; the $+0.74$ pp k-NN reversal at epoch 600 is observed across both $N_{\text{local}} = 6$ seeds. The VTAB-1k random-init floor (Section 5.2) is single-seed, and per-task floor variance under different inits is uncharacterised. The VTAB-1k $\times$ N_{local} evaluation at 32K $\times$ 600 epochs (Section 5.5) uses two paired seeds for $N_{\text{local}} = 6$ and one for $N_{\text{local}} = 0$ and $N_{\text{local}} = 8$. Any single-seed gap within ± 2 pp is reported as an ordering signal, not a confirmed magnitude.
- **Resolution.** All experiments use 64×64 input images. The token-budget argument in Section 6.2 predicts that the multi-crop effect should weaken at higher resolutions (e.g., 96×96 or 128×128), where local crops carry more unique source content. We did not test this prediction.
- **Backbone.** We use ViT-Tiny/8 throughout. Larger backbones such as ViT-Small or ViT-Base may show different N_{local}–scale interactions and are untested in our setting.
- **Downstream scope.** Evaluation covers CIFAR-10 linear probe, Tiny-ImageNet weighted k-NN, and VTAB-1k linear probe across 19 tasks. Object detection and semantic segmentation are not evaluated. Whether the N_{local}–task-category pattern observed on VTAB-1k extends to dense prediction tasks is an open question.
- **Compute scope and hyperparameters.** We use DINO [1] reference defaults throughout, without hyperparameter search; interactions between N_{local} and learning-rate, weight-decay, or EMA momentum schedules are unexplored. Per-step compute also differs across N_{local} values, so we report epoch-budget rather than compute-budget comparisons. The k-NN protocol uses $k = 20$ and $\tau = 0.07$ throughout; sensitivity to these choices at 64×64 is uncharacterised.
- **`diabetic_retinopathy` artefact.** All N_{local} values and pretraining sizes score exactly 73.59% on this task — the majority-class rate, indicating the frozen ViT-Tiny CLS token carries no discriminative signal for retinal fundus images at 64×64. We exclude it from per-category interpretation but keep it in the overall mean for VTAB-1k reporting consistency.

8 Conclusion

We studied DINO [1] pretraining at sub-ImageNet scale across three axes: pretraining-set size, multi-crop view budget (N_{local}), and training duration.

Representation quality grows monotonically with pretraining-set size under a uniform 200-epoch budget, approaching the supervised CIFAR-10 reference at the largest split while offering no benefit at the smallest. The canonical $N_{\text{local}} = 6$ recipe is not always the best choice: removing local crops ($N_{\text{local}} = 0$) matches or beats it at medium splits under tight compute, and its lead narrows — and partially reverses on k-NN — with extended training. Increasing past six local crops gives no further benefit. On downstream transfer, the best N_{local} depends on task category: removing local crops favours natural-image tasks, keeping them favours specialized tasks, and structured tasks are largely indifferent.

The published view that multi-crop universally helps DINO [1] — established at ImageNet scale — does not transfer cleanly to sub-ImageNet pretraining at 64×64. We recommend $N_{\text{local}} = 0$ with 200 epochs for compute-constrained natural-image targets, and $N_{\text{local}} = 6$ or $N_{\text{local}} = 8$ with 600+ epochs for specialized targets. At sub-ImageNet scale, DINO's multi-crop recipe is not one-size-fits-all — the right choice emerges from training budget and downstream task category rather than from the ImageNet default.

9 Responsible Research

This section addresses four concerns: the implementation and reproducibility setup, alignment of the released artefacts with the FAIR principles (Findable, Accessible, Interoperable, Reusable), data ethics of the pretraining and evaluation datasets, and the scope within which our conclusions hold.

9.1 Implementation and Reproducibility

All runs execute on a single NVIDIA A100-40GB using PyTorch [16; 17] with HuggingFace Transformers and Kornia [18] and bfloat16 autocast. We hand-ported the DINO [1] components from scratch and verified the port against the reference `facebookresearch/dino` implementation with six numerical tests (ViT forward, DINO loss, EMA teacher update, centering-buffer update, teacher-temperature warmup, and multi-crop augmentation tensors), all matching to within 10^{-5}. All random sources are seeded; the main data-efficiency curve and the 32K extended-training $N_{\text{local}} = 0$ run use three paired seeds (data seeds $\{42, 43, 44\}$, model seeds $\{0, 1, 2\}$), with single-seed runs covering the remaining N_{local} conditions. Source code, per-run configs, pretrained checkpoints, and result CSVs are released [19], so every figure regenerates from the CSVs without re-running pretraining.

9.2 FAIR principles

The FAIR principles call for research outputs to be Findable, Accessible, Interoperable, and Reusable [20]. *Findable and Accessible:* the code and result files are released publicly [19], downloadable without registration. *Interoperable:* we use standard formats throughout — Python source,

YAML configuration files, plain CSV results with documented columns, and PyTorch checkpoint files — so no proprietary tooling is needed to read them. *Reusable:* each run is accompanied by its configuration and seeds (Section 9.1), and the released CSVs are sufficient to regenerate every figure of Section 5 without re-running pretraining.

9.3 Data ethics

We pretrain on Tiny-ImageNet [4] and evaluate on CIFAR-10 [21] and the Visual Task Adaptation Benchmark (VTAB-1k) [15]. Tiny-ImageNet and CIFAR-10 are public research benchmarks with permissive licences and are widely used in the visual representation learning literature. Recent audits document biases and consent issues in the person-related ImageNet classes [22]. The Tiny-ImageNet class list is dominated by everyday objects, animals, and scenes, and does not include the person-subtree categories audited in [22]. VTAB-1k aggregates 19 sub-datasets with mixed licences; we use the canonical `train800val200` split per task and apply no per-image filtering. One VTAB-1k task, `diabetic_retinopathy`, contains medical fundus images sourced from a Kaggle competition (EyePACS) with subject consent; we use this task only under its public-release terms and disclose its degenerate behaviour at this resolution in Section 7. Both pretraining and evaluation datasets are English-labelled and Western-centric, which limits the inclusivity of conclusions drawn from them.

9.4 Use of AI tools

We used large language models as writing and coding assistants during the project. The authors verified all experiments, claims, and final text against original sources and the experimental data; no claim in this paper was generated without independent verification.

9.5 Scope of conclusions

We restrict our claims to the recipe and protocol we measured. The results apply to DINO [1] pretraining with the ViT-Tiny/8 backbone at 64×64 input on Tiny-ImageNet subsets between 1K and 100K images. We evaluate by CIFAR-10 linear probe, Tiny-ImageNet weighted k-NN, and VTAB-1k linear transfer. The N_{local} sweep covers $\{0, 4, 6\}$ at the uniform 200-epoch protocol and $\{0, 6, 8\}$ at the 32K $\times$ 600-epoch extended-training sub-protocol. We did not test 224×224 input, larger backbones such as ViT-S or ViT-B, or the full 65,536-dimensional projection head. The $N_{\text{local}} = 0$ result at the 32K split under the 200-epoch protocol and the k-NN ordering change at 600 epochs are properties of this regime and should not be read as universal claims about DINO at other resolutions or scales. Pretraining the eight splits across the multi-seed main curve, the 200-epoch N_{local} ablation, and the 32K $\times$ 600-epoch extended training consumed roughly 500 A100-GPU-hours; we report this following [23; 24]. We discourage deploying a DINO checkpoint pretrained on $\leq$ 1K images in safety-critical settings, since we observe no measurable benefit over random initialisation at that scale.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[4] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," CS 231N course report, Stanford University, 2015.

[5] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, and S. Belongie, "When does contrastive visual representation learning work?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2024.

[8] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jégou, and E. Grave, "Are large-scale datasets necessary for self-supervised pre-training?," *arXiv preprint arXiv:2112.10740*, 2021.

[9] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[11] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[13] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[14] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[15] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby, "A large-scale study of representation learning with the visual task adaptation benchmark," *arXiv preprint arXiv:1910.04867*, 2019.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[17] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, *et al.*, "PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation," in *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2024.

[18] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: an open source differentiable computer vision library for PyTorch," in *Winter Conference on Applications of Computer Vision*, 2020.

[19] L. Margulis, "Code and result files for "data-efficiency of self-supervised learning with DINO multi-crop"." https://github.com/Leonid-Margulis-03/RP, 2026.

[20] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, 2016.

[21] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., University of Toronto, 2009.

[22] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[23] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[24] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021.