
Data Density Scaling Laws for Image Self-Distillation

Alvard Barseghyan Ani Vanyan Hakob Tamazyan Hrant Khachatryan
YerevaNN and Yerevan State University, Yerevan, Armenia
{alla, ani, hakob, hrant}@yerevann.com

Abstract

Current scaling law formulations suggest that increasing the number of unique samples in a dataset improves model performance, while repeated exposure to the same samples provides diminishing value. In this work, we examine this assumption for image self-distillation models. Through systematic experiments across diverse datasets, namely Maxar, SATLAS, Waymo, nuPlan, Ego4D, ImageNet, and YFCC15M, we demonstrate that the effects of sample repetition on downstream performance are non-trivial and dataset-dependent. We identify settings where repeatedly training on a smaller random subset can outperform training on a larger set of unique samples. This finding highlights a limitation of existing scaling laws. To address this gap, we propose a modified version of a Chinchilla-style scaling law that explicitly accounts for two factors: the number of unique samples and repetition per sample. This formulation provides a framework for understanding data–compute trade-offs and improves prediction in some regimes where repetition plays a central role, although its benefits are not consistent across datasets. Finally, we examine how subset selection and augmentation affect downstream performance.

1 Introduction

Recent progress in foundation models has been strongly associated with scaling both model size and dataset size (1; 2; 3). In language modeling, empirical scaling laws suggest that performance improves predictably as a function of model parameters, compute, and the number of unique training tokens (4; 3). Similar trends have motivated large-scale representation learning in computer vision (5; 6; 7; 8), where increasing the number of training images is generally assumed to improve representation quality.

However, as foundation models expand into domain-specific visual settings such as remote sensing and autonomous driving, the assumption that increasing the number of unique samples always improves performance becomes less reliable. Many real-world visual corpora contain substantial semantic redundancy, near-duplicate samples, and uneven informational density. In such settings, increasing nominal dataset size may not correspond to increasing effective semantic coverage.

In this work, we examine where existing data-constrained scaling laws are insufficient to describe the trade-off between unique samples and repetition in visual self-distillation. Through experiments on Maxar satellite imagery (9), SATLAS (10), the Waymo Open Dataset (11), nuPlan, Ego4D, ImageNet (12), and YFCC15M, we demonstrate that the interaction between unique sample count and repetition density is highly dataset-dependent.

Most strikingly, we find that a random subset containing only 1.5% of the Full Maxar pool improves average downstream kNN performance by 3.11 percentage points compared to pretraining on the full 50M-image pool under an identical budget of 200M image presentations. Rather than rejecting scaling laws altogether, our findings suggest that their assumptions do not fully capture the observed smaller-subset advantage.

To study this phenomenon, we formalize three core variables:

1. **Unique data size (U):** The number of distinct samples in the selected pretraining pool.
2. **Repetitions or Density (R):** The average number of exposures per encountered unique sample during pretraining.
3. **Compute Budget (C):** The total number of source-image presentations, used as a compute proxy within a fixed model and training recipe.

To account for partial passes, we define encountered support as $U_{\text{seen}} = \min(C, U)$ and repetition as $R = C/U_{\text{seen}}$, giving $C = U_{\text{seen}}R$. We use the terms *repetition*, *density*, and *repetition density* interchangeably for R throughout the paper.

Compute-only scaling laws collapse unique support and repetition into the single exposure term C . Data-constrained laws already distinguish these quantities, but their assumptions about the value of repeated data do not capture every observed trade-off in visual self-distillation.

While prior work in NLP (13) shows that repeated exposure to the same tokens leads to diminishing returns, we observe different behavior in some visual self-distillation settings. Unlike autoregressive language modeling, repeated exposures in visual self-distillation are not equivalent to replaying identical training targets. Each image exposure is transformed through stochastic augmentations, while the teacher network evolves continuously through EMA updates. As a result, repeated views of the same image correspond to dynamically changing supervisory signals rather than static token repetition. These differences motivate our study, but do not establish the mechanism behind the subset advantage.

Our contributions are threefold. First, we provide empirical evidence that repeatedly training on random subsets can outperform training on substantially larger sets of unique samples under fixed compute budgets. Second, we compare existing scaling laws with a density-interaction extension of a Chinchilla-style compute-only law that captures interactions between unique images and repetitions. This formulation improves held-out prediction on some datasets, although its benefits are not consistent across all seven. Finally, we examine how the subset advantage depends on model size, self-distillation method, subset selection, and augmentation.

Our results suggest that scaling behavior in image self-distillation depends not only on dataset size, but also on how the training budget is allocated between unique support and repeated exposure.

2 Related Work

Self-Supervised Pretraining and Distillation: The success of large-scale pretraining in NLP (1; 2) has motivated the development of vision foundation models trained using self-supervised learning (5; 6; 7; 8). Existing approaches are broadly divided into Masked Image Modeling (MIM) and self-distillation methods. MIM approaches, such as MAE (5), reconstruct masked image regions or latent tokens, while self-distillation approaches learn representations by matching a student network to the predictions of a teacher network (14). Our work focuses on iBOT (15), which combines masked prediction with online self-distillation through an EMA-updated teacher. This training paradigm is particularly relevant to our study because repeated exposures to the same image do not correspond to identical supervisory signals.

Unlike autoregressive language modeling, repeated exposures in visual self-distillation are not equivalent to replaying identical training targets. Each exposure corresponds to a different stochastic augmentation, while the teacher network evolves continuously through EMA updates. As a result, repeated views of the same image provide dynamically changing supervisory signals rather than static token reuse. This distinction may partially explain why repetition behaves differently in visual SSL compared to data-constrained language modeling.

Scaling Laws and Compute-Optimal Training: Scaling laws have emerged as a central framework for understanding the relationship between model performance, dataset size, and compute (4; 3). In language modeling, compute-optimal training laws suggest that performance depends jointly on model size and the number of unique training tokens, while repeated exposure to the same data rapidly yields diminishing returns (13). Similar assumptions are commonly adopted in visual representation learning, where larger image collections are generally expected to improve downstream performance.

Recent work has also begun exploring scaling behavior under data filtering and constrained compute budgets (16; 17; 18). These studies suggest that data quality, filtering strategy, and subset selection interact strongly with training compute, and that curated subsets can outperform substantially larger unfiltered corpora in certain regimes. However, existing formulations still primarily treat total data exposure as a single scaling axis.

Our work revisits this assumption in the context of visual self-distillation. Traditional scaling laws implicitly collapse unique sample count and repetition into a single exposure term, treating repeated exposures as approximately interchangeable with additional unique data under fixed compute budgets. In contrast, our results suggest that repetition density and unique sample count play fundamentally different roles in highly redundant visual domains.

Data Curation, Filtering, and Pruning: Modern large-scale pretraining increasingly relies on data curation and filtering to improve training efficiency and dataset quality (7; 19; 20). Prior work has shown that carefully selected subsets can outperform substantially larger unfiltered datasets, particularly when large-scale corpora contain duplicates, noisy samples, or low-information examples (21). Recent studies further demonstrate that the utility of curated subsets depends strongly on the available compute budget (16).

Our work is closely related to this line of research, but differs in its primary objective. Unlike prior work on data pruning or filtering, our goal is not solely to identify smaller high-quality subsets, but to study how repetition density and unique sample count interact under fixed compute budgets. In particular, we show that scaling behavior itself changes across datasets with different redundancy structures. While ImageNet largely follows conventional scaling behavior, highly redundant domains such as Maxar exhibit regimes in which repeatedly training on curated subsets can outperform training on substantially larger collections of unique samples.

Redundancy, Duplicate Data, and Effective Dataset Size: Several recent works argue that nominal dataset size can substantially overestimate the effective information content of a corpus due to duplicates, semantic overlap, and low-information samples (20; 21). These findings suggest that representation quality depends not only on dataset scale, but also on the distribution and diversity of samples.

Our work extends this perspective to self-distillation learning by analyzing how semantic redundancy interacts with repetition during training. We hypothesize that performance degradation in large, highly redundant datasets does not arise solely from the presence of low-information or noisy samples, but also from their *joint interaction within a fixed subset* (22). In particular, certain samples may become mutually detrimental when co-occurring in the same training set under self-distillation objectives, leading to amplified redundancy effects and less effective representation learning under fixed compute budgets. Sampling can partially mitigate this effect by decorrelating such harmful co-occurrences, thereby reducing their collective negative impact on optimization. To quantify effective semantic coverage, we measure Vendi scores (23) over DINOv2 embeddings and analyze how diversity structure influences downstream performance across different repetition regimes.

3 Experimental Design

3.1 Pretraining Datasets

We pretrain independently on seven datasets covering four visual domains.

Remote sensing: Maxar and SATLAS contain satellite imagery, with Full experimental pools of 50,000,000 and 10,969,092 images, respectively. For Maxar, the original images were divided into non-overlapping 500×500 -pixel tiles, producing 51,197,237 tiles; our Full training pool uses 50M of these.

Driving: Waymo and nuPlan contain 5,806,392 and 25,481,542 images, respectively. Both provide driving scenes with temporal redundancy from nearby video frames.

Egocentric video: Ego4D provides 38,387,515 frames depicting everyday activities from a first-person perspective.

Natural images: ImageNet contains 1,281,167 curated, object-centered training images, while our YFCC15M pool contains 14,825,236 images drawn from YFCC.

Throughout the paper, “Full” refers to the complete experimental training pool, which may differ from the complete source dataset.



Figure 1: Samples from pretraining datasets

3.2 Sampling Strategies

The experimental design serves two purposes. First, we vary the unique-data support U using uniformly sampled random subsets to study its interaction with repetition at fixed training budgets. We define encountered support as $U_{\text{seen}} = \min(C, U)$ and repetition as $R = C/U_{\text{seen}}$. Second, we compare different subset-selection methods at matched subset sizes and training budgets to examine whether the observed subset advantage depends on the selection strategy.

Random sampling selects images uniformly from the parent dataset without explicitly favoring image quality or semantic coverage. The main scaling-law experiments use only eligible random subsets and Full. For Maxar, these contain 50M, 25M, 12.5M, 8.045397M, 2M, 0.75M, and 0.4M images. Audited unstable checkpoints and the ImageNet random subsets containing 1,241,131 and 1,201,094 images are excluded from both fitting and evaluation at every budget and seed.

Hierarchical sampling constructs subsets using representation-space clusters. We extract [CLS] embeddings with a pretrained DINOv2 model (7), then apply hierarchical k -means clustering across four levels with $k \in \{50k, 10k, 5k, 1k\}$. Balanced sampling across clusters reduces the dominance of densely populated clusters.

We also compare random and hierarchical sampling with DASSOT, a graph-matching approach to balanced data selection (24), and quality-based selection using the Natural Image Quality Evaluator (NIQE) (25). This comparison uses 0.75M-image Maxar subsets at matched training budgets and is separate from scaling-law fitting.

3.3 Model Pretraining

We pretrain iBOT with a ViT-B/16 backbone following the standard protocol (15), using the same multi-resolution augmentations as the original implementation. Optimization is performed with AdamW using a per-GPU batch size of 200 across 8 GPUs, resulting in a global batch size of $B = 1600$. All experiments are run in distributed mode with FlashAttention-2 (26) for efficiency.

We perform pretraining across compute budgets $C \in \{10M, 20M, 30M, 40M, 50M, 100M, 200M, 300M, 400M\}$. Figure 2 illustrates the experimental grid and evaluates whether subset selection meaningfully affects pretrained representation quality. We use two sampling strategies: random uniform sampling and hierarchical sampling.

Given batch size $B = 1600$, the number of optimization steps is $S = C/B$ (e.g., 10M exposures $\approx 6,250$ steps, 50M $\approx 31,250$ steps).

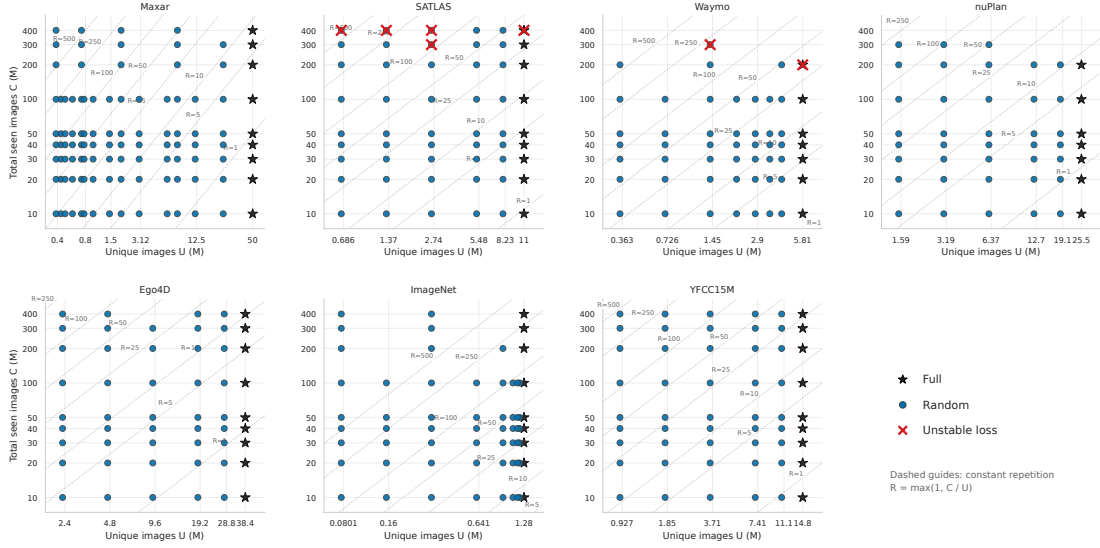


Figure 2: Pretraining grid across datasets. Each point corresponds to a run with unique data size U and compute budget C (both in millions). Colors denote Maxar, Waymo, and ImageNet. Markers indicate sampling strategy: full data, random subsets, and hierarchical subsets. Diagonal lines show constant repetition R , highlighting the trade-off between unique data and reuse. The fire icons indicate non-stable pre-trainings.

3.4 Downstream Evaluation

We evaluate frozen representations using kNN classification and logistic regression on normalized final-layer [CLS] embeddings. For each pretraining group and classification task, we select $k \in \{1, 5, 10, 30, 50\}$ using mean accuracy across Full-50M reference seeds, breaking ties toward the smallest k . This choice remains fixed across models, budgets, and seeds. Logistic regression uses the Cyanure solver with the saved task-level regularization selections.

Remote sensing: Maxar and SATLAS share six classification tasks: Brick Kiln, EuroSAT, RESISC45, UCMerced, PatternNet, and AID.

Driving: Waymo and nuPlan share BDD100K scene and weather classification and ACDC condition classification. Their kNN aggregates additionally include BDD100K MOT instance- and category-level scores. These tracking scores have no k and are excluded from logistic-regression aggregates.

General and egocentric recognition: Ego4D, ImageNet, and YFCC15M share 24 classification tasks, including ImageNet-1K validation in both evaluation suites. Ego4D additionally uses two EPIC-KITCHENS-100 noun-recognition tasks, based on single-frame and five-frame pooled features.

We average available task scores within each pretraining seed, then across available seeds, keeping kNN and logistic regression separate. Missing scores are not replaced by zero. The main results use kNN; logistic-regression results are reported in Appendix D. Task details and fixed k values appear in Appendix B.

4 Empirical Observations

Figure 3 summarizes the empirical data-density behavior under kNN evaluation; the corresponding logistic-regression curves are provided in Appendix C. A central empirical question is whether data-constrained scaling behavior from language modeling transfers to image self-distillation. Data-constrained language-model laws (13) assign diminishing value to repeated samples and generally favor larger unique support at fixed compute. However, this trade-off need not describe every visual self-distillation setting, where each repeated image is transformed by stochastic augmentations and supervised by an evolving EMA teacher. We examine this question across seven pretraining datasets

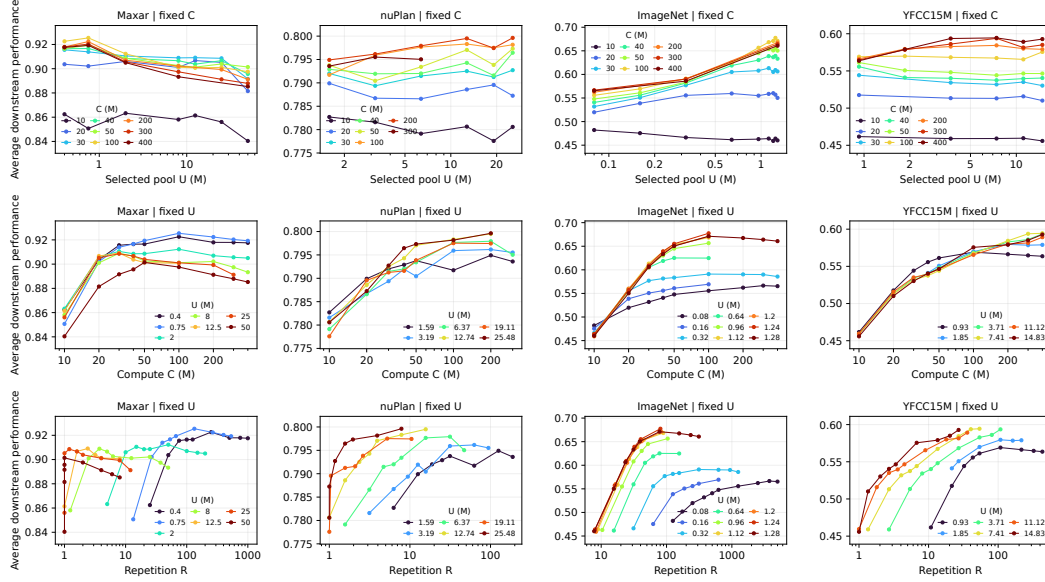


Figure 3: Behaviour of different random subsets under the kNN evaluation protocol. The first column fixes compute C and plots downstream performance as a function of unique images U . The middle column fixes U and plots performance as a function of compute C . The third column fixes U and plots performance as a function of repetition R . The figure shows the trade-off between unique support and repetition for full-data and random-subset runs.

using eligible random-subset and Full observations. We call the highest-scoring observed support at a given budget, including Full, the *best evaluated subset*.

Observation 1: the fixed-compute U/R trade-off is dataset-dependent. At fixed compute, the datasets show qualitatively different behavior. ImageNet is support-limited at high compute: at $C = 100\text{M}$, Full reaches 0.6743 average kNN performance, while the 640,584-image subset, approximately half of Full, reaches 0.6278. The driving datasets show smaller differences between many supports. For example, the best evaluated Waymo subset exceeds Full by only 0.20 percentage points at $C = 100\text{M}$, while Full is the best evaluated nuPlan support at $C = 200\text{M}$.

Maxar reverses the usual intuition: substantially reduced support can outperform Full at matched compute. At $C = 200\text{M}$, the random 0.75M-image subset exceeds Full by 3.11 percentage points. Smaller subsets also outperform Full on SATLAS, Ego4D, and YFCC15M, with best-evaluated gains of 0.65, 0.85, and 0.54 percentage points, respectively. Thus, the fixed-compute trade-off is not a universal curve: reducing unique support can substantially hurt, have a small effect, or improve downstream performance.

Observation 2: the trade-off changes with compute. The relationship among C , U , and R is not fixed even within a dataset. ImageNet is less support-limited at lower compute: at $C = 20\text{M}$, the 640,584-image subset reaches 0.5613, compared with 0.5520 for Full. This ordering reverses at $C = 100\text{M}$. On Ego4D, Full is the best evaluated support at 50M presentations, whereas the 2,399,219-image subset leads at 100M and 200M. On YFCC15M, the best evaluated support changes from 926,577 images at 50M to Full at 100M and 7,412,618 images at 200M.

Maxar shows a more persistent smaller-subset advantage: its random 0.75M-image subset outperforms Full at 50M, 100M, 200M, and 400M presentations. These differences indicate that repetition cannot be treated as a dataset-independent correction to compute. Its observed value depends jointly on the domain, selected support, and training budget, making extrapolation from low-budget runs a separate challenge.

Observation 3: smaller random subsets can outperform Full. Maxar is the clearest case. At matched compute, Full training underperforms much smaller random subsets that are repeated many

times. At $C = 100\text{M}$, the random 0.75M-image subset reaches 0.9251 average kNN performance, compared with 0.8976 for the 50M-image Full pool. At $C = 400\text{M}$, the corresponding values are 0.9210 and 0.8853. Thus, a subset containing only 1.5% of Full improves performance by 2.76 and 3.57 percentage points at these budgets.

This behavior shows that increasing unique support is not always beneficial under a fixed training recipe and budget, even without specialized subset curation. One possible explanation is that smaller subsets reduce task-irrelevant variation or conflicting modes. However, random subsampling also changes coverage and co-occurrence, so these comparisons do not establish that particular images are harmful or that repetition alone causes the advantage.

Observation 4: repetition can help, saturate, or hurt. At fixed support, additional presentations increase repetition once the first pass is complete. Small Maxar subsets improve strongly at lower budgets and then largely saturate, while Full Maxar decreases from 0.9015 at $C = 50\text{M}$ to 0.8853 at $C = 400\text{M}$. ImageNet Full initially improves from 0.6527 at 50M to 0.6743 at 100M, then declines to 0.6626 at 400M. In contrast, Ego4D’s 2,399,219-image subset improves from 0.4969 at 50M to 0.5126 at 400M. Additional exposure therefore has different effects depending on the dataset and support size. These conclusions are tied to kNN evaluation; the separate logistic-regression curves show how the observed behavior depends on the downstream protocol.

We exclude audited unstable checkpoints before making these comparisons. In particular, Waymo has no eligible Full-200M measurement, so its gain over Full is undefined at that budget. The reported differences are descriptive point estimates and do not by themselves establish statistical significance.

Together, these observations motivate a density-aware view of image self-distillation. The empirical object is not compute alone or unique data alone, but their interaction through encountered support $U_{\text{seen}} = \min(C, U)$ and repetition $R = C/U_{\text{seen}}$. Existing data-constrained laws provide useful baselines, while the observed smaller-subset advantages motivate testing more flexible support-repetition interactions.

5 Density-Aware Scaling Law

The empirical observations above raise two questions: Can existing scaling-law parameterizations explain the interaction among total compute, unique data, and repetition in image self-distillation? And, by using only small-budget runs, can scaling laws determine which pretraining subset to use at a target compute budget to maximize average downstream accuracy?

Fitting protocol. We fit every law independently for each pretraining dataset. The target is average downstream accuracy across the corresponding frozen-representation evaluation tasks, so the reported error measures how well a law predicts transfer performance rather than pretraining loss. Unless stated otherwise, the main results use the kNN aggregate. Parameters are optimized only on low-compute runs with $C \leq 50\text{M}$ image views and evaluated on higher-compute runs with $C > 50\text{M}$. All laws use the same underlying quantities: total seen images C , selected pool size U , encountered unique images $U_{\text{seen}} = \min(C, U)$, induced repetition $R = C/U_{\text{seen}}$. We measure C and U_{seen} in millions of image presentations and unique images, respectively. Repetition R is dimensionless. We use the same post-spike exclusions described in Section 4. We report mean absolute error (MAE) separately in the fit range and in the extrapolation range.

Existing laws. We compare three baseline parameterizations adapted to predict downstream accuracy at fixed model size. The first baseline is a compute-only Chinchilla-style power law,

$$\hat{y}(C) = L - \frac{A}{C^\alpha}, \quad (1)$$

which captures improvement with total training exposure but cannot distinguish two runs with the same C and different (U, R) . The second baseline uses a data-constrained effective-data parametrization:

$$U_{\text{eff}}(U_{\text{seen}}, R) = U_{\text{seen}} \left[1 + \tau \left(1 - \exp \left(-\frac{R-1}{\tau} \right) \right) \right], \quad \hat{y}(U, R) = L - \frac{A}{U_{\text{eff}}(U_{\text{seen}}, R)^\beta}, \quad (2)$$

Here τ controls the saturation scale of the contribution from additional passes. This law gives additional passes diminishing value, with the rate controlled by τ . The third baseline is an adaptation

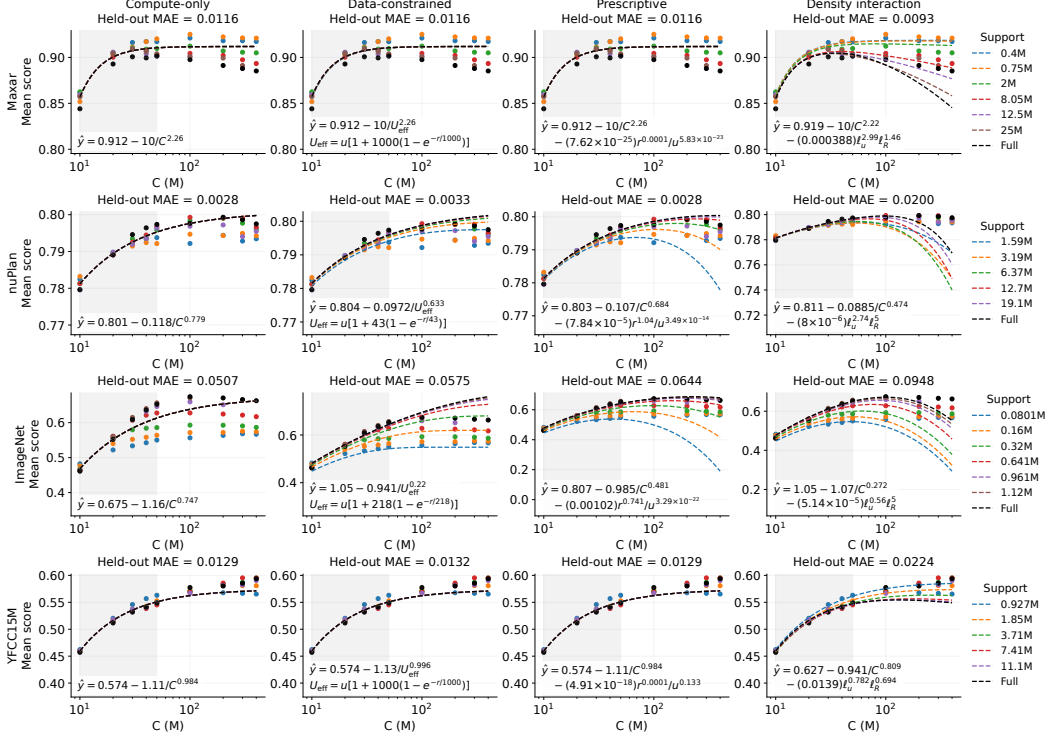


Figure 4: Scaling-law fits and extrapolations under the kNN evaluation protocol. Columns compare three existing-law baselines—compute-only, data-constrained, and prescriptive—with our density-interaction law. These four forms test whether compute alone, diminishing returns from repetition, a repetition penalty, or a support–repetition interaction captures the observed behavior. Rows show Maxar, nuPlan, ImageNet, and YFCC15M. For compactness, SATLAS and Ego4D are omitted because they show similar qualitative patterns to Maxar, while Waymo shows similar observed trends to nuPlan. Curves are fitted only on the shaded low-compute region and evaluated on higher-compute runs. Fitting curves for all datasets and the logistic-regression comparison can be found in Appendix ??.

of the prescriptive law, which adds a repetition-dependent penalty:

$$\hat{y}(C, U, R) = L - \frac{A}{C^\alpha} - \frac{B(R-1)^\delta}{U_{\text{seen}}^\eta}, \quad (3)$$

The compute-only law gives the same prediction to all subsets at a fixed compute budget. The data-constrained law allows repeated exposure to help, but cannot predict declining performance at fixed unique support. The prescriptive law can model such a decline. However, under our parameter constraints, none of these baselines can predict that a smaller subset strictly outperforms a larger one at the same compute budget. Their fitted predictions are also similar on several datasets; Appendix ?? examines these similarities.

Density interaction. To capture these differences, we extend the compute-only baseline with a joint correction for unique support and repetition.

$$\text{Perf}(C, U, R) = L - \frac{A}{C^\alpha} - D[\log(1 + U_{\text{seen}})]^\beta [\log(1 + R)]^\gamma. \quad (4)$$

The compute term captures improvement with total training exposure. The interaction term adjusts the prediction according to how this exposure is divided between unique images and repetitions. We interpret it as a phenomenological density penalty, not a first-principles law. One possible mechanism is that harmful, low-information, or domain-conflicting modes are scattered across the unique sample

Table 1: kNN scaling-law MAE in accuracy units, fitted on $C \leq 50\text{M}$ and evaluated on $C > 50\text{M}$. Only eligible random subsets and Full are included. Each task uses a fixed k chosen from Full-50M reference seeds. Bold and underlining indicate the lowest and second-lowest distinct unrounded held-out errors in each dataset.

Law	Maxar		SATLAS		Waymo		nuPlan		Ego4D		ImageNet		YFCC15M	
	extrap.	fit	extrap.	fit	extrap.	fit	extrap.	fit	extrap.	fit	extrap.	fit	extrap.	fit
Compute-only	0.0050	0.0116	0.0022	0.0060	0.0019	0.0021	0.0010	<u>0.0028</u>	0.0012	<u>0.0033</u>	0.0234	0.0507	0.0033	0.0129
Data-constrained	0.0050	0.0116	0.0022	0.0058	0.0017	0.0024	0.0009	0.0033	0.0012	0.0033	0.0125	<u>0.0575</u>	0.0034	0.0132
Prescriptive	0.0050	<u>0.0116</u>	0.0020	<u>0.0057</u>	0.0015	0.0023	0.0009	0.0028	0.0011	0.0034	0.0110	0.0644	0.0033	<u>0.0129</u>
Density interaction	0.0032	0.0093	0.0016	0.0043	0.0015	<u>0.0022</u>	0.0008	0.0200	0.0012	0.0027	0.0080	0.0948	0.0019	0.0224

pool. A smaller random subset may omit some of these modes, while repeated training may amplify the effects of those retained. This remains a hypothesis rather than a mechanism established by our experiments. The logarithms avoid treating nominal unique images or repetitions as linearly independent units, and the exponents β and γ control dataset-specific sensitivity to unique support and repetition density. These coefficients should be interpreted jointly with D , not as standalone mechanistic measurements.

Main results. Table 1 shows that density interaction achieves the lowest held-out MAE on Maxar, SATLAS, and Ego4D, and remains close to the best baseline on Waymo. However, it performs worse on nuPlan, ImageNet, and YFCC15M. Thus, allowing a smaller-subset advantage improves prediction in some settings but does not guarantee better extrapolation.

We next examine whether the observed subset advantage persists across model sizes, self-distillation methods, subset-selection strategies, and augmentation settings.

Effect of model size. To test whether the subset advantage is specific to ViT-B, we also train iBOT with ViT-S and ViT-L on Maxar. For both additional model sizes, we evaluate four dataset sizes at training budgets from 10M to 200M image presentations. At 100M presentations, the random 0.75M-image subset achieves average kNN accuracy of 0.9207 with ViT-S and 0.9196 with ViT-L, compared with 0.8994 and 0.8796 for full-data training. The subset advantage also appears under logistic-regression evaluation. These results show that the finding persists across the three tested model sizes.

Effect of the self-distillation method. To test whether the subset advantage is specific to iBOT, we also train DINOv2 with ViT-B on Maxar. At 100M image presentations, the random 0.75M-image subset achieves average kNN accuracy of 0.9182, compared with 0.9095 for full-data training. This advantage persists through 300M presentations and also appears under logistic-regression evaluation. These results support the finding across two self-distillation methods, while its generalization to other self-supervised learning methods remains untested.

Effect of subset selection. To test whether the subset advantage depends on the selection method, we compare random sampling, hierarchical selection, DASSOT, and NIQE-based quality selection on Maxar. At 50M image presentations, the 0.75M-image subsets achieve average kNN accuracies of 0.9202, 0.9187, 0.9202, and 0.9216, respectively, compared with 0.9015 for full-data training. The advantage across all four methods shows that it does not require a particular selection strategy. However, these comparisons do not fully separate the effects of repetition, redundancy, image quality, and semantic coverage.

Effect of augmentations. We hypothesize that unique images and repeated exposure may contribute differently to learning in visual self-distillation. We hypothesize that repeated exposure remains useful beyond the additional variation provided by strong augmentation. To test this, we train iBOT on a random 0.75M-image Maxar subset with reduced augmentation. We disable color jitter, grayscale, Gaussian blur, and solarization, while retaining cropping, horizontal flipping, normalization, and masking.

Reduced augmentation lowers average kNN accuracy at 100M presentations from 0.9251 to 0.8910. However, accuracy increases from 0.8400 at 10M presentations to 0.8969 at 400M. This supports the hypothesis that useful repetition does not depend entirely on the removed augmentations. The retained crops and masks still introduce variation, and without a reduced-augmentation Full control, we cannot determine whether the subset advantage also persists.

The statistical tests show that the accuracy differences captured by the Maxar scaling-law analysis are not small numerical artifacts: 124 of 187 matched pairwise comparisons are significant at $p < 0.05$, while the fitted law achieves low train and heldout error. This supports the paper’s main outcome that, in high-redundancy remote-sensing data, scaling is governed not only by compute and unique-image count but also by data density/repetition, with carefully chosen reduced-U subsets often matching or exceeding full-data training at the same compute.

Conclusion

We studied empirical scaling behavior for image self-distillation when total compute C , unique support U , and repetition $R = C/U$ are varied independently. The main finding is that data density is not a minor implementation detail: reduced unique support can hurt early on ImageNet, remain nearly neutral on Waymo, and substantially improve Maxar at matched compute. Existing compute-only and data-constrained laws do not consistently capture these behaviors, while a density-interaction law gives a better empirical description in the tested extrapolation range.

Acknowledgements

We thank Evan Shelhamer for useful comments. This research was supported by the HESC MESCS RA grant No. 24RL-1B049.

References

- [1] Radford, A., J. Wu, R. Child, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [2] Brown, T. B., B. Mann, N. Ryder, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.
- [3] Hoffmann, J., S. Borgeaud, A. Mensch, et al. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.
- [4] Kaplan, J., S. McCandlish, T. Henighan, et al. Scaling Laws for Neural Language Models. *arXiv e-prints*, arXiv:2001.08361, 2020.
- [5] He, K., X. Chen, S. Xie, et al. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022.
- [6] Wang, W., H. Bao, L. Dong, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186. 2023.
- [7] Oquab, M., T. Darcet, T. Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.
- [8] Siméoni, O., H. V. Vo, M. Seitzer, et al. DINOv3, 2025.
- [9] MaxarTechnologies. Open data program. <https://registry.opendata.aws/maxar-open-data>, 2022. Accessed: 2022 Apr 1.
- [10] Bastani, F., P. Wolters, R. Gupta, et al. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 16726–16736. IEEE, 2023.

- [11] Sun, P., H. Kretzschmar, X. Dotiwalla, et al. Scalability in perception for autonomous driving: Waymo open dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [12] Deng, J., W. Dong, R. Socher, et al. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.
- [13] Muennighoff, N., A. M. Rush, B. Barak, et al. Scaling data-constrained language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 2023.
- [14] Caron, M., H. Touvron, I. Misra, et al. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE, 2021.
- [15] Zhou, J., C. Wei, H. Wang, et al. ibot: Image BERT pre-training with online tokenizer. *CoRR*, abs/2111.07832, 2021.
- [16] Goyal, S., P. Maini, Z. C. Lipton, et al. Scaling laws for data filtering - data curation cannot be compute agnostic. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22702–22711. IEEE, 2024.
- [17] Wiegand, G.-H., L. Raichle, R. Staedeli, et al. Is more data worth the cost? dataset scaling laws in a tiny attention-only decoder, 2026. DATA-FM Workshop @ ICLR 2026.
- [18] Ye, J., V. Feldman, K. Talwar. Cram less to fit more: Training data pruning improves memorization of facts, 2026. DATA-FM Workshop @ ICLR 2026.
- [19] Gadre, S. Y., G. Ilharco, A. Fang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*. 2023.
- [20] Abbas, A., K. Tirumala, D. Simig, et al. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *CoRR*, abs/2303.09540, 2023.
- [21] Sorscher, B., R. Geirhos, S. Shekhar, et al. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, vol. 35, pages 19523–19536. 2022.
- [22] Hu, Y., P. Hu, H. Zhao, et al. Most influential subset selection: Challenges, promises, and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37. 2024.
- [23] Friedman, D., A. B. Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.
- [24] Van Assel, H., R. Balestrieri. A graph matching approach to balanced data sub-sampling for self-supervised learning. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*.
- [25] Mittal, A., R. Soundararajan, A. C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.
- [26] Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- [27] Lee, J., N. R. Brooks, F. Tajwar, et al. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17):e2018863118, 2021.
- [28] Helber, P., B. Bischke, A. Dengel, et al. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019.
- [29] Cheng, G., J. Han, X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

- [30] Yang, Y., S. D. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In D. Agrawal, P. Zhang, A. E. Abbadi, M. F. Mokbel, eds., *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*, pages 270–279. ACM, 2010.
- [31] Zhou, W., S. Newsam, C. Li, et al. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018.
- [32] Xia, G.-S., J. Hu, F. Hu, et al. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.
- [33] Yu, F., H. Chen, X. Wang, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645. 2020.
- [34] Sakaridis, C., D. Dai, L. Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10765–10775. 2021.
- [35] Krause, J., M. Stark, J. Deng, et al. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561. 2013.
- [36] Krizhevsky, A. Learning multiple layers of features from tiny images. Tech. rep., University of Toronto, 2009.
- [37] Wah, C., S. Branson, P. Welinder, et al. The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.
- [38] Cimpoi, M., S. Maji, I. Kokkinos, et al. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613. 2014.
- [39] Goodfellow, I. J., D. Erhan, P. L. Carrier, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.
- [40] Maji, S., J. Kannala, E. Rahtu, et al. Fine-grained visual classification of aircraft. Tech. rep., 2013.
- [41] Nilsback, M.-E., A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*. 2008.
- [42] Bossard, L., M. Guillaumin, L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*. 2014.
- [43] Stallkamp, J., M. Schlipsing, J. Salmen, et al. The german traffic sign recognition benchmark: A multi-class classification competition. In *Proceedings of the 2011 International Joint Conference on Neural Networks*, pages 1453–1460. 2011.
- [44] LeCun, Y., L. Bottou, Y. Bengio, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [45] Veeling, B. S., J. Linmans, J. Winkens, et al. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 210–218. 2018.
- [46] Parkhi, O. M., A. Vedaldi, A. Zisserman, et al. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.
- [47] Coates, A., H. Lee, A. Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223. 2011.

- [48] Fei-Fei, L., R. Fergus, P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.
- [49] Lomonaco, V., D. Maltoni. CORE50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, vol. 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR, 2017.
- [50] Zhu, C., F. Xiao, A. Alvarado, et al. EgoObjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20110–20120. 2023.
- [51] Quattoni, A., A. Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. 2009.
- [52] Damen, D., H. Doughty, G. M. Farinella, et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

A Pretraining Dataset Vendi Scores

Table 2: Vendi scores by dataset, sampling strategy, and target subset percentage.

Dataset	Sampling	100prc	75prc	50prc	37.5prc	25prc	12.5prc	6.25prc
Maxar	Random	47.99	47.99	47.99	47.90	47.90	47.96	47.89
	Hierarchical	–	60.51	60.51	66.38	66.38	70.61	74.67
Waymo	Random	73.38	73.40	73.39	73.42	73.40	73.48	73.42
	Hierarchical	–	87.81	107.26	119.67	136.68	–	172.22
ImageNet	Random	589.65	589.75	589.59	589.61	589.73	589.77	589.64
	Hierarchical	–	584.31	587.46	–	–	–	–

Note: Maxar uses the nearest available subset size to each target percentage based on 50M training images. Waymo and ImageNet use exact available percentage matches; missing entries are shown as –.

B Downstream Tasks

Maxar and SATLAS are evaluated on six remote-sensing classification tasks (Table 3). Waymo and nuPlan use three driving classification tasks, with two additional BDD100K tracking scores included in their kNN aggregates (Table 4). Ego4D, ImageNet, and YFCC15M share 24 classification tasks; Ego4D additionally uses EPIC-KITCHENS-100 noun recognition with single-frame and five-frame pooled features (Table 5).

For each pretraining group and classification task, $k \in \{1, 5, 10, 30, 50\}$ maximizes mean accuracy across Full-50M reference seeds; ties select the smallest k . This choice remains fixed across models, budgets, and seeds. Tracking scores have no k . We aggregate kNN and logistic-regression results separately, averaging available task scores within each seed and then across seeds.

Table 3: Remote-sensing downstream tasks and fixed reference k values.

Task	Train	Eval	Classes	Maxar k	SATLAS k
Brick Kiln (27)	58,890	7,361	2	10	5
EuroSAT (28)	16,200	5,400	10	5	5
RESISC45 (29)	18,900	6,300	45	1	5
UCMerced (30)	1,260	420	21	1	1
PatternNet (31)	18,240	6,080	38	1	5
AID (32)	6,000	2,000	30	1	1

Table 4: Driving downstream tasks and fixed reference k values. Tracking scores enter only the kNN aggregate; their sample counts are unspecified.

Task	Train	Eval	Classes	Waymo k	nuPlan k
BDD100K scene (33)	69,863	10,000	7	30	30
BDD100K weather (33)	69,863	10,000	7	30	30
ACDC condition (34)	1,600	2,000	4	50	30
BDD100K MOT instance	–	–	–	–	–
BDD100K MOT category	–	–	–	–	–

Table 5: General and egocentric downstream tasks. Fixed reference k values are shown for Ego4D (Ego), ImageNet (IN), and YFCC15M (YFCC). A dash indicates that the task is not evaluated for that group.

Task	Train	Eval	Classes	Ego k	IN k	YFCC k
ImageNet-1K (12)	1,281,167	50,000	1,000	30	10	30
Stanford Cars (35)	8,144	8,041	196	1	1	1
CIFAR-100 (36)	50,000	10,000	100	10	10	30
CIFAR-10 (36)	50,000	10,000	10	10	10	10
CUB-200 (37)	5,994	5,794	200	30	30	30
DTD (38)	1,880	1,880	47	10	1	10
FER-2013 (39)	28,709	7,178	7	1	1	1
FGVC-Aircraft (40)	3,334	3,333	100	1	1	1
Flowers-102 (41)	1,020	1,020	102	1	1	1
Food-101 (42)	75,750	25,250	101	50	30	50
GTSRB (43)	26,640	12,630	43	50	30	50
MNIST (44)	60,000	10,000	10	10	5	10
PCAM (45)	262,144	32,768	2	50	50	50
Oxford-IIIT Pets (46)	3,680	3,669	37	10	30	30
STL-10 (47)	5,000	8,000	10	10	5	10
Caltech-101 (48)	3,030	5,647	101	10	10	1
RESISC45 (29)	18,900	6,300	45	1	5	10
UCMerced (30)	1,260	420	21	1	1	1
EuroSAT (28)	16,200	5,400	10	1	5	5
Brick Kiln (27)	58,890	7,361	2	5	5	5
CORE50 identity (49)	119,894	44,972	50	10	50	50
EgoObjects category (50)	447,735	164,240	638	50	50	50
EgoObjects identity (50)	27,690	124,956	3,510	1	1	1
MIT Indoor-67 (51)	5,360	1,340	67	30	10	10
EPIC noun (frame) (52)	67,217	9,668	300	10	–	–
EPIC noun (pool) (52)	67,217	9,668	300	5	–	–

EgoObjects counts refer to object crops. Category classification uses a 638-label vocabulary, with 352 categories observed in training and 275 in evaluation. Identity classification uses capture-disjoint reference and query crops within the official test subset; Train denotes the reference bank. EPIC-KITCHENS-100 counts refer to action segments, with evaluation on public validation. Its 300-label vocabulary contains 289 labels observed in training; all validation segments are retained.

C Logistic-regression density curves

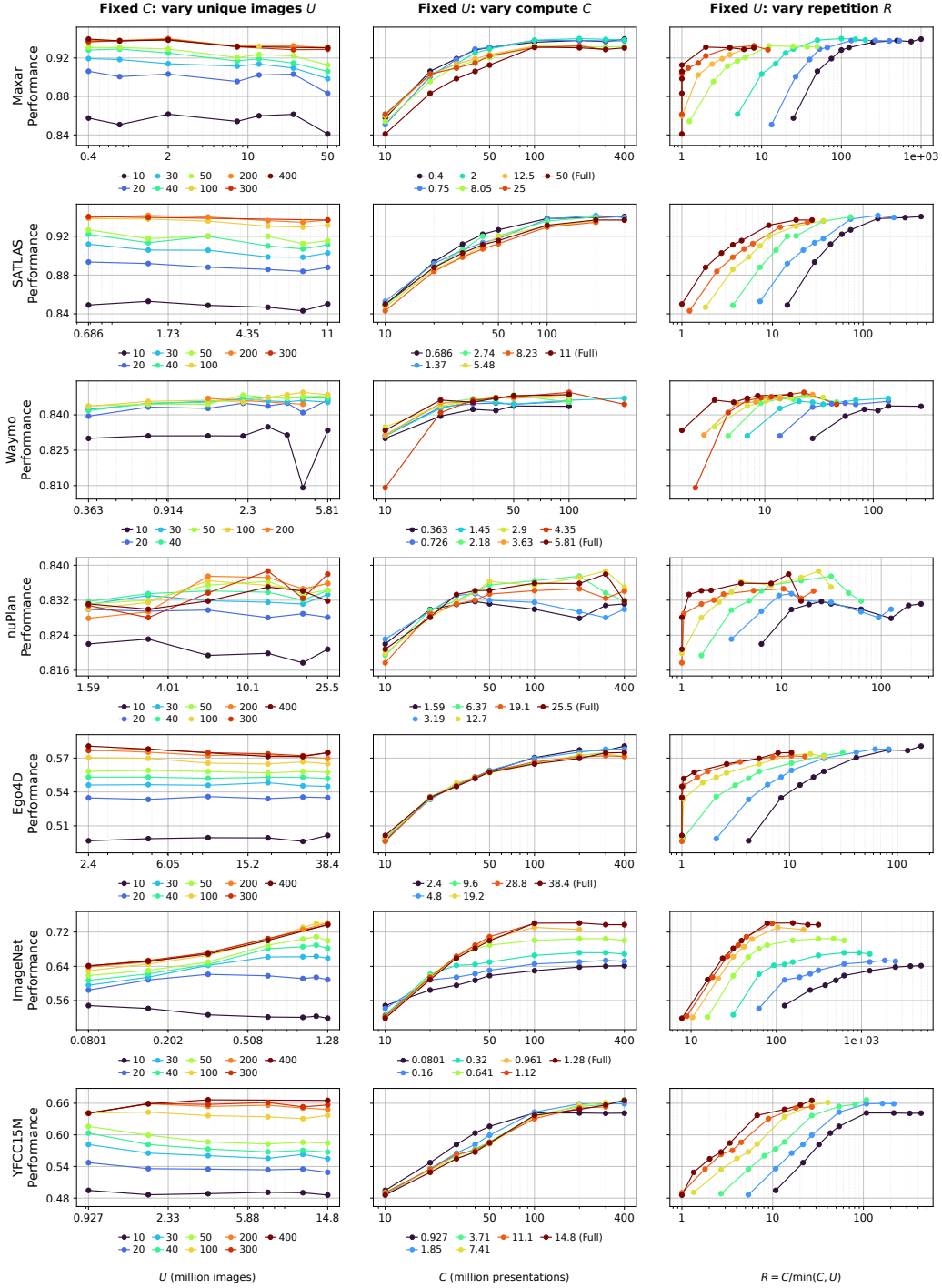


Figure 5: Unique-support and repetition trade-offs under logistic-regression evaluation. Rows correspond to Maxar, SATLAS, Waymo, nuPlan, Ego4D, ImageNet, and YFCC15M. Columns show mean downstream accuracy as unique images U vary at fixed compute C , compute varies at fixed U , and repetition $R = C / \min(C, U)$ varies at fixed U , respectively. Points represent random-subset and Full-data runs; lines connect observations. Colors indicate compute in the left column and unique support in the middle and right columns, which share a color key. Horizontal axes are logarithmic; image counts are in millions.

D Logistic-regression scaling-law results

E k-NN density curves

F k-NN scaling-law results

Table 6: Logistic-regression scaling-law comparison. Each law is fitted on runs with $C \leq 50\text{M}$ and evaluated on runs with $C > 50\text{M}$. Lower MAE is better. We use $\ell_U = \log(1 + U)$ and $\ell_R = \log(1 + R)$.

Law	Formula	Maxar fit	Maxar extrap.	Waymo fit	Waymo extrap.	ImageNet fit	ImageNet extrap.
Compute-only	$L - A/C^\alpha$	0.0068	0.0089	0.0034	0.0018	0.0166	0.0249
Data-constrained	$L - A/U^{\beta_{\text{eff}}}$	0.0119	0.0089	0.0034	0.0067	0.0140	0.0283
Density interaction	$L - A/C^\alpha - D\ell_U^\beta \ell_R^\gamma$	0.0053	0.0070	0.0029	0.0037	0.0114	0.0112
Unique-only density	$L - A/C^\alpha - D_U \ell_U^\beta$	0.0053	0.0070	0.0030	0.0018	0.0142	0.0190
Repeat-only density	$L - A/C^\alpha - D_R \ell_R^\gamma$	0.0068	0.0089	0.0029	0.0036	0.0122	0.0145
Separate density	$L - A/C^\alpha - D_U \ell_U^\beta - D_R \ell_R^\gamma$	0.0053	0.0070	0.0028	0.0062	0.0119	0.0134
Raw- U interaction	$L - A/C^\alpha - D U^\beta \ell_R^\gamma$	0.0053	0.0071	0.0029	0.0038	0.0114	0.0117
Raw- R interaction	$L - A/C^\alpha - D \ell_U^\beta R^\gamma$	0.0053	0.0070	0.0029	0.0044	0.0114	0.0298

G Scaling-law ablation plots

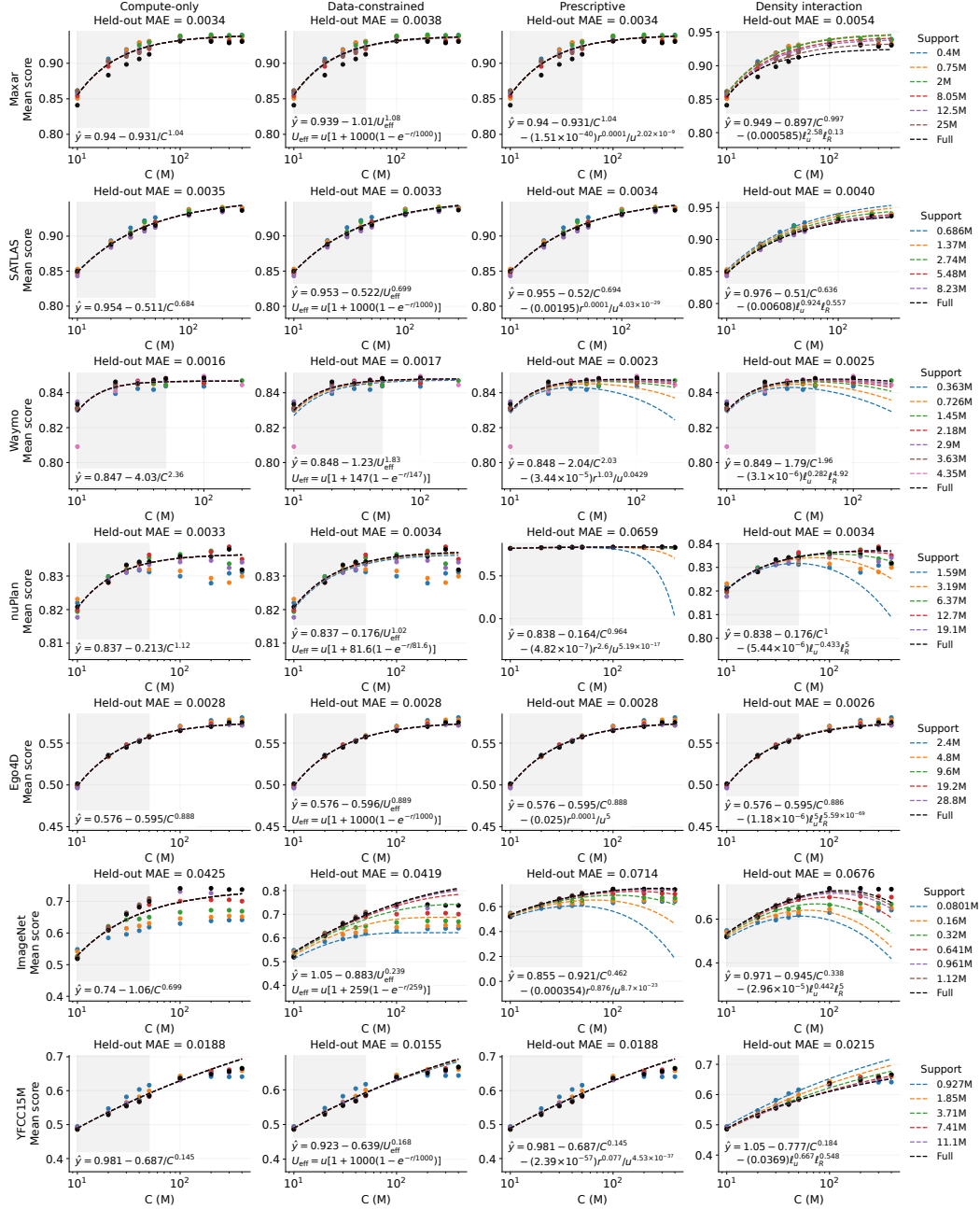


Figure 6: Scaling-law fits and extrapolations under the logistic-regression evaluation protocol for all seven datasets. Columns compare the compute-only Chinchilla-style law, the data-constrained effective-data law, the prescriptive scaling law, and our density-interaction law. Rows correspond to Maxar, SATLAS, Waymo, nuPlan, Ego4D, ImageNet, and YFCC15M. Curves are fitted only in the shaded low-compute region ($C \leq 50M$) and evaluated on higher-compute runs ($C > 50M$). Each panel shows the fitted equation and held-out MAE in accuracy units. Parameters are fitted separately from the kNN evaluation.

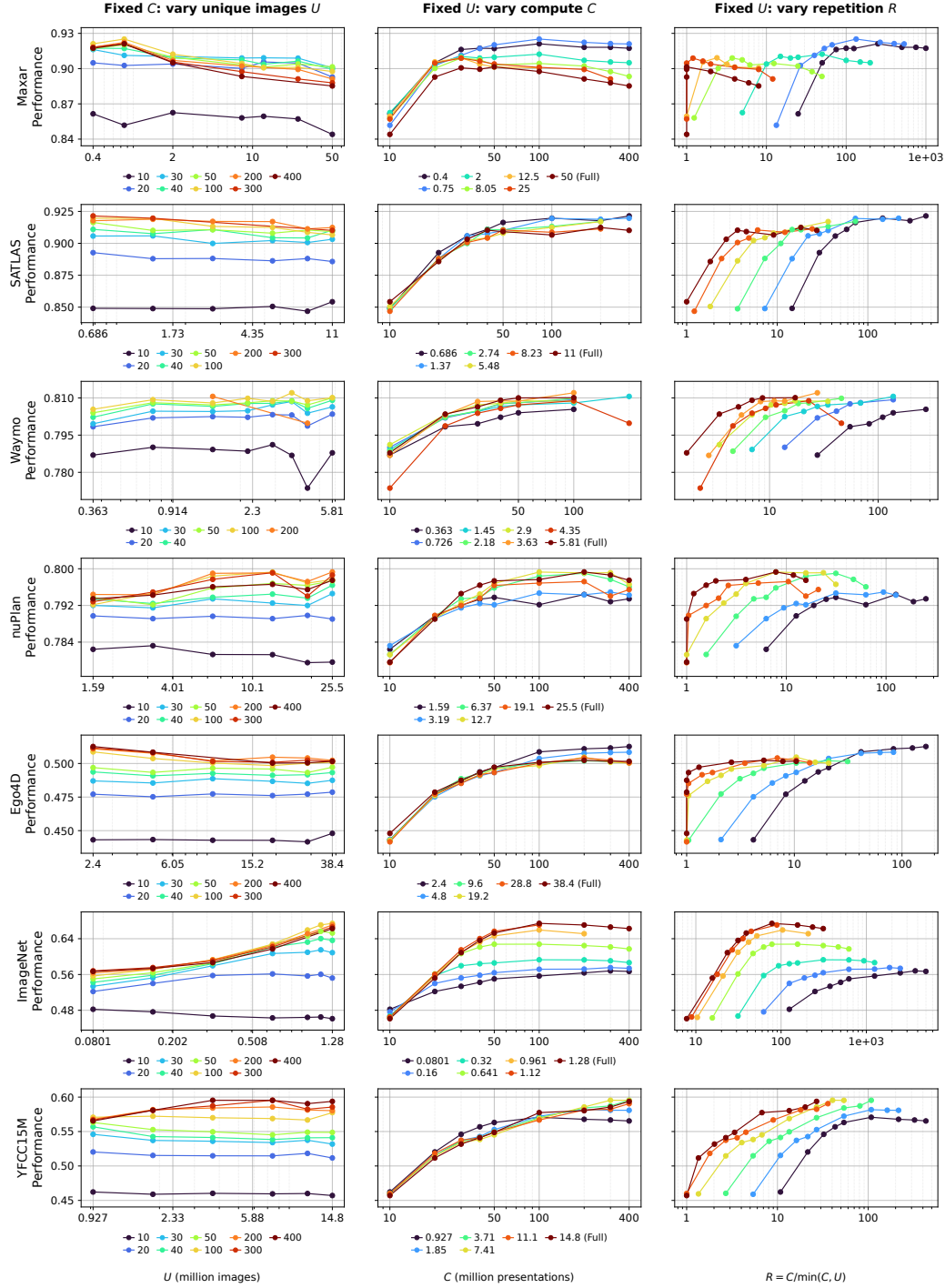


Figure 7: Unique-support and repetition trade-offs under kNN evaluation. Rows correspond to Maxar, SATLAS, Waymo, nuPlan, Ego4D, ImageNet, and YFCC15M. The left column fixes compute C and varies unique images U ; the middle column fixes U and varies C ; the right column fixes U and varies repetition $R = C/\min(C, U)$. Points show mean downstream accuracy for random-subset and Full-data runs, with lines connecting observations. Colors indicate compute in the left column and unique support in the middle and right columns, which share a color key. For each pretraining group and downstream task, k is selected from the Full-data 50M reference runs and held fixed across evaluated models. Horizontal axes are logarithmic; image counts are in millions.

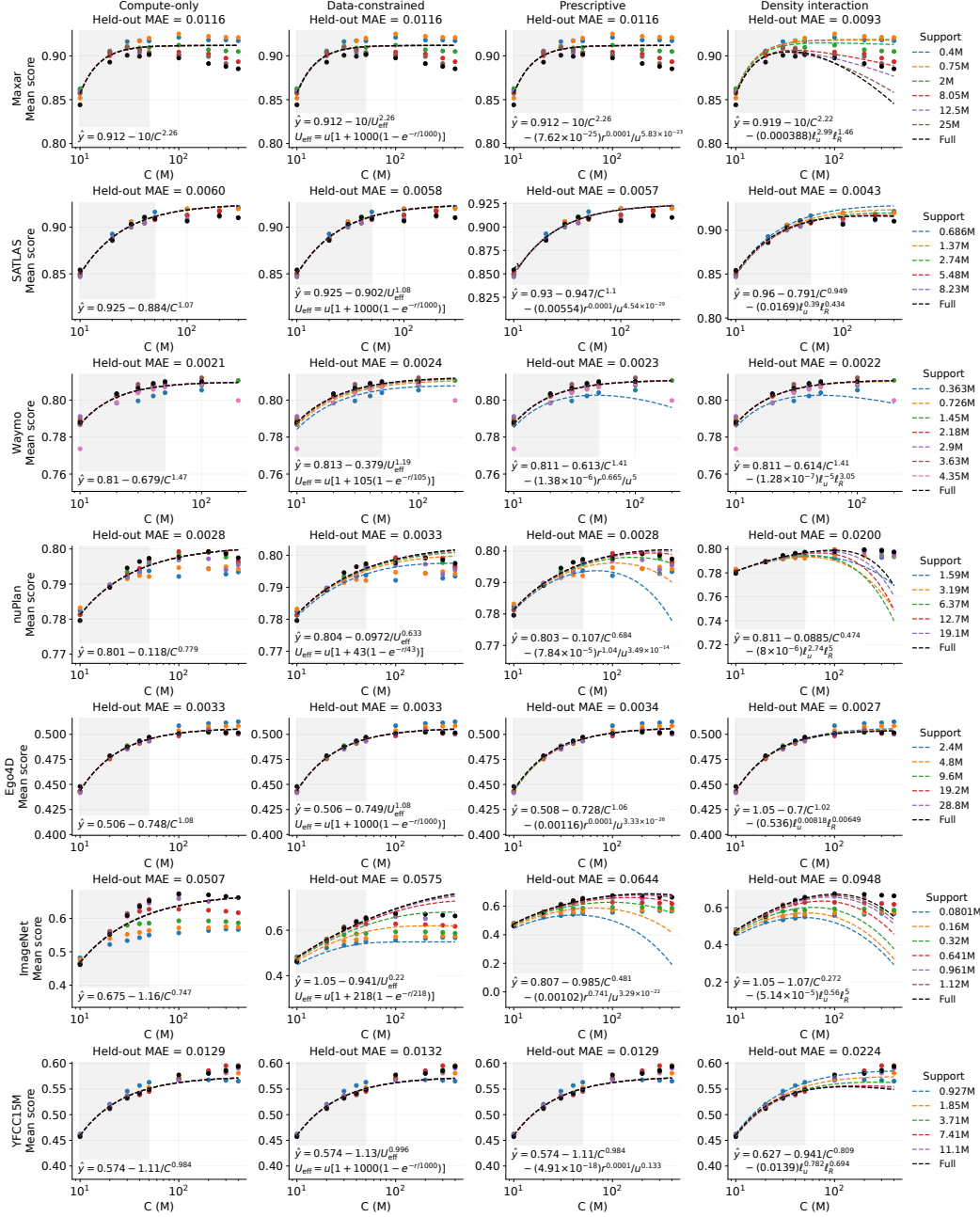


Figure 8: Scaling-law fits and extrapolations under the kNN evaluation protocol for all seven datasets. Columns compare the compute-only Chinchilla-style law, the data-constrained effective-data law, the prescriptive scaling law, and our density-interaction law. Rows correspond to Maxar, SATLAS, Waymo, nuPlan, Ego4D, ImageNet, and YFCC15M. Curves are fitted only in the shaded low-compute region ($C \leq 50M$) and evaluated on higher-compute runs ($C > 50M$). Each panel shows the fitted equation and held-out MAE in accuracy units.

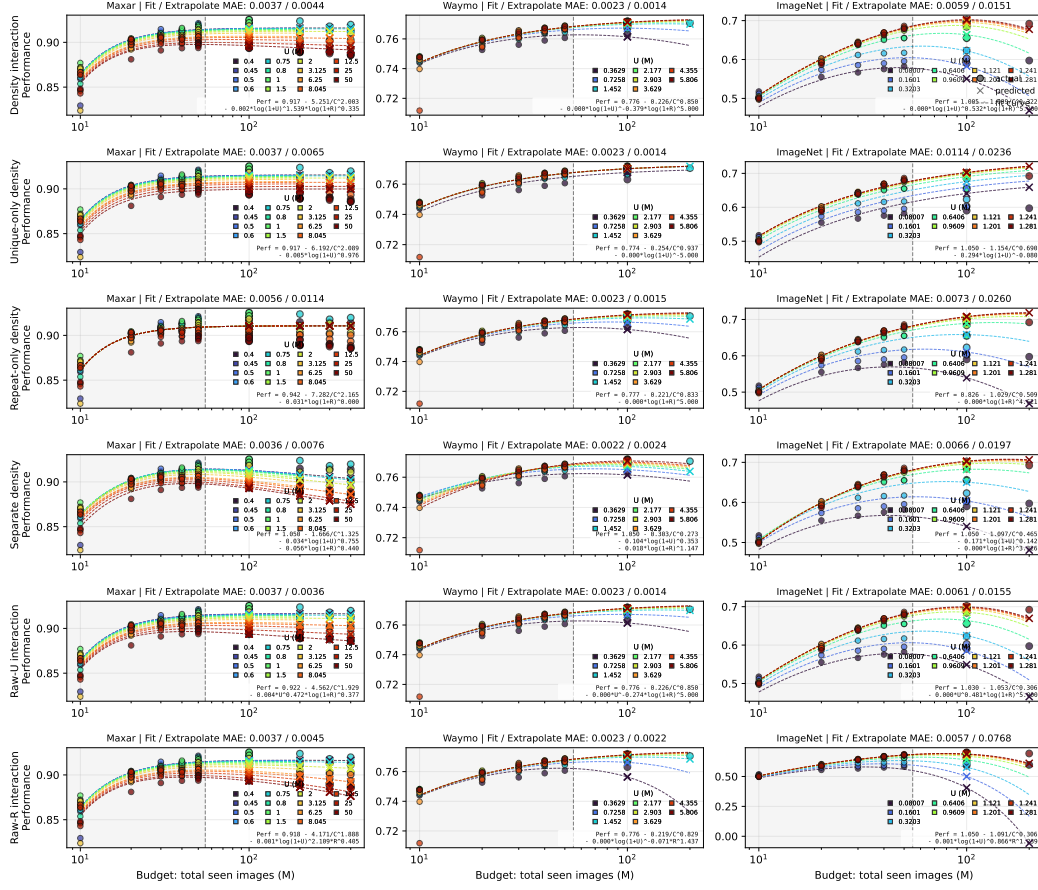


Figure 9: kNN ablation plots for the density-interaction law. Rows compare the proposed interaction law against unique-only, repeat-only, separate-additive, raw- U , and raw- R variants.

H Fitted density-interaction coefficients

Table 7: Fitted coefficients for the density-interaction law, $\text{Perf} = L - A/C^\alpha - D[\log(1+U)]^\beta[\log(1+R)]^\gamma$. Coefficients are fitted independently per dataset and evaluation protocol.

Eval.	Dataset	L	A	α	D	β	γ
kNN	Maxar	0.9169	5.2508	2.0030	0.002364	1.5387	0.3352
kNN	Waymo	0.7760	0.2263	0.8500	0.000001	-0.3792	5.0000
kNN	ImageNet	1.0052	1.0393	0.3219	0.000046	0.5322	5.0000
Logres	Maxar	0.9550	0.5118	0.6884	0.000984	2.2653	0.0000
Logres	Waymo	0.7829	0.7126	1.3312	0.000004	0.0899	5.0000
Logres	ImageNet	1.0500	0.9501	0.3084	0.000034	0.5158	5.0000

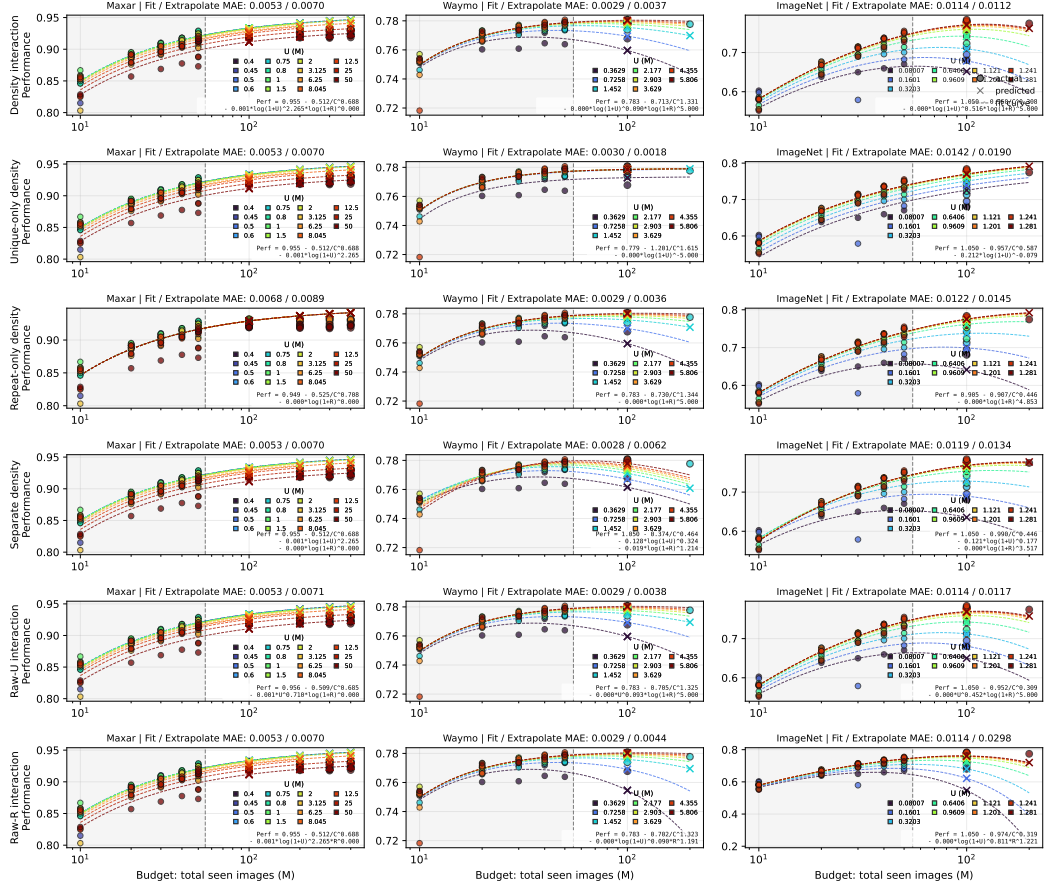


Figure 10: Logistic-regression ablation plots for the density-interaction law.

I Subset Selection

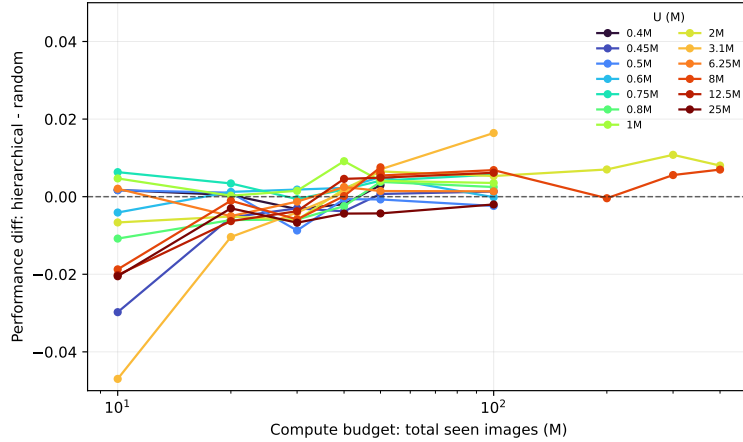


Figure 11: Maxar matched-pair comparison between hierarchical-clustering selection and random selection. Each line fixes the subset size U and plots the downstream-performance difference, hierarchical minus random, as compute increases.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The main claims in the abstract and introduction are directly supported by controlled experiments analyzing dataset size, repetition, and diversity effects under fixed compute budgets using self-supervised pretraining.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss limitations including use of a single backbone (ViT-B/16), reliance on iBOT-style pretraining, and evaluation limited to standard benchmarks and frozen-feature evaluation protocols.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The work is entirely empirical and does not include theoretical results or proofs.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide full details on architectures, training protocol (iBOT, AdamW, batch size, distributed setup), dataset construction (random and hierarchical sampling), and evaluation procedures (kNN and linear probing), enabling reproduction of results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used are publicly available (ImageNet-1K, Waymo Open Dataset, Maxar-derived datasets, and standard benchmarks). Experimental procedures are fully specified and code can be released to reproduce results.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all relevant training and evaluation details including optimizer, batch size, compute budgets, sampling strategies, evaluation protocols, and hyperparameter selection.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report averaged results over multiple seeds.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We specify global batch size, distributed training setup, and compute budgets for all experiments, allowing reproduction and estimation of computational cost per run.

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research follows the NeurIPS Code of Ethics, uses publicly available datasets, and does not involve human subjects or sensitive data collection.

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive impacts in improving understanding of data efficiency in foundation models, and potential negative impacts related to compute scaling and accessibility of large-scale training.

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not release generative models or high-risk systems requiring safeguards.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets are properly cited with their sources and known licenses, including ImageNet-1K, Waymo Open Dataset, and standard vision benchmarks.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We introduce structured subsets constructed via random and hierarchical sampling over public datasets, fully described and reproducible.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[N/A\]](#)

Justification: The work does not involve human subjects or crowdsourcing.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[N/A\]](#)

Justification: No human subject research is conducted.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [\[N/A\]](#)

Justification: Large language models are not part of the methodology or experimental pipeline.