BEYOND DATA SIZE: EXPLORING THE IMPACT OF DATASET DIVERSITY AND DENSITY IN SELF-DISTILLATION LEARNING

Alvard Barseghyan[1,2]**,Ani Vanyan**[1,2]**, Hakob Tamazyan**[1,2]**,**
Hrant Khachatrian[1,2]
[1] YerevaNN
[2] Yerevan State University
{alla,ani,hakob,hrant}@yerevann.com

ABSTRACT

Current scaling laws suggest that maximizing unique data is key to superior pretraining. For self-distillation models like iBOT, we show that data density (repetition) and data diversity (as measured by Vendi score) can be as critical as data size (the total number of unique samples). Wide range of experiments on a large remote sensing dataset demonstrate that seeing a smaller, high-quality subset multiple times outperforms a single pass over a massive stream of unique samples under equivalent compute. Based on these results, we propose a predictive scaling law that models downstream performance as a joint function of unique data size, data density and data diversity. We demonstrate the extrapolation power of the proposed formula.

1 INTRODUCTION

The emergence of foundation models in computer vision has been heavily influenced by task-agnostic representation learning in Natural Language Processing (NLP) (Radford et al., 2019; Raffel et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022). In the NLP domain, large-scale pretraining on raw internet text has enabled models to achieve remarkable performance across a variety of tasks in both zero-shot and few-shot settings (Brown et al., 2020). This success has motivated the development of vision-based foundation models (Bommasani et al., 2021), which aim to learn transferable representations applicable to both image-level tasks, such as classification, and pixel-level tasks, such as segmentation.

Established scaling laws in NLP (Kaplan et al., 2020; Hoffmann et al., 2022) suggest that model performance scales predictably with increased data volume. However, our findings challenge this "more-is-better" dogma within the context of self-distillation learning. We demonstrate that superior performance can be achieved not by simply expanding the dataset, but by strategically removing samples from the original pool to optimize its composition. To investigate this phenomenon, we pose two fundamental questions:

- Is the initial ***Data Diversity*** sufficient to support high-quality representation learning?
- Is ***Data Density*** (redundancy) a necessary component for stability during self-distillation learning?

In this work we have three concepts: **1. Data Size (U)** - the total number of unique samples, **2. Data Density (R)** - repetition of each unique sample during pretraining, **3. Data Diversity (V)** - quality metric across unique samples which is measured by Vendi score (Friedman & Dieng, 2023; Pasarkar & Dieng, 2024).

We conduct our primary experiments on the Maxar satellite imagery dataset (MaxarTechnologies, 2022), which provides a unique empirical setting for studying data redundancy. Due to repeated temporal imaging of the same geographic locations under varying seasonal and atmospheric condi-

tions, the dataset contains substantial near-duplicate content. This high-density characteristic makes Maxar ideal for investigating compute-data tradeoffs in self-supervised pretraining.

To construct curated subsets at various scales, we first extract representation vectors for the entire collection using a pretrained DINOv2 model (Oquab et al., 2024). We then employ a clustering-based curation method to form balanced subsets at multiple target sizes (Vo et al., 2024). The impact of these filtered versions is evaluated through k-NN classification across four diverse remote sensing benchmarks.

Our experimental results demonstrate that "more is not always better": training on a strategically pruned subset containing only 1.5% of the total data can improve average downstream k-NN performance by approximately 3 percentage points relative to training on the full dataset.

Building upon these concepts, we propose two candidates of **predictive scaling laws** that attempt to model the downstream performance of self-supervised models as a joint function of U, R, and V. Surprisingly, the one inspired by data-constrained scaling laws did not fit well our data. Instead, a more straightforward extension of Chinchilla law worked better. Our findings align with emerging trends in large-scale pretraining; for instance, recent advancements in the DINO family (Siméoni et al., 2025) emphasize the necessity of data curation and the use of homogeneous batches.

2 Related Work

Self-Supervised Pretraining: Following the significant achievements of self-supervised pretraining on large-scale, web-crawled data in natural language processing, similar paradigms (He et al., 2022; Wang et al., 2023; Zhou et al., 2021; Oquab et al., 2024; Siméoni et al., 2025) have gained prominence for natural images. Masked Image Modeling (MIM) methods, such as MAE, learn representations by reconstructing missing content from masked inputs (He et al., 2022). In parallel, self-distillation approaches learn semantic invariances by matching a teacher network's predictions to a student's, typically with the teacher updated via an Exponential Moving Average (EMA) of the student weights (Zhou et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). These self-distillation-based algorithms have demonstrated superior performance on diverse downstream tasks; notably, recent large-scale models such as DINOv2 and DINOv3 exhibit exceptional transferability and retrieval-style capabilities without the need for task-specific fine-tuning.

In this work, we employ iBOT (Zhou et al., 2021) as our pretraining framework. iBOT is particularly well-suited for our study as it unifies masked token prediction with self-distillation via an online tokenizer and EMA teacher updates, consistently producing representations that are highly effective for k-NN-based evaluation.

Data Curation: Constructing high-quality subsets from large, unlabeled image collections is a significant challenge, particularly in the presence of extreme redundancy and long-tail distributions. Modern pretraining pipelines increasingly rely on sophisticated curation to remove near-duplicates and enhance dataset balance. For instance, the DINOv2 pipeline incorporates a dedicated copy-detection stage to eliminate near-duplicate images during dataset construction (Oquab et al., 2024). Similarly, SemDeDup leverages pretrained embeddings to identify semantic duplicates, filtering them to maximize data efficiency (Abbas et al., 2023).

For balanced subset selection, Vo et al. (2024) proposed a clustering-based approach utilizing hierarchical k-means followed by balanced sampling; this enables the construction of subsets at specific target scales while preserving overall diversity. Alternatively, Van Assel & Balestriero formulate subset selection as a graph-matching problem over pairwise similarities, which provides an advantage when clusters are not well-separated in the embedding space. Crucially, recent studies emphasize that curation should not be treated as compute-agnostic. The utility of any fixed subset is inherently tied to its training duration (i.e., the number of repetitions), which motivates the need for compute-aware filtering strategies (Tirumala et al., 2022). In this work, we adopt the hierarchical clustering-based curation approach (Vo et al., 2024) to construct balanced subsets of Maxar imagery across multiple scales, allowing us to explicitly study the interplay between curation and repetition.

Scaling laws: In language modeling, empirical scaling laws characterize the functional relationship between loss, model capacity, and training compute, providing a framework for compute-optimal resource allocation (Kaplan et al., 2020). Subsequent Chinchilla-style analyses revealed that many

large language models (LLMs) were significantly under-trained relative to their parameter counts, leading to revised compute-optimal tradeoffs between model size and data volume (Hoffmann et al., 2022). Particularly relevant to redundancy-heavy domains is the study of data-constrained scaling, which examines the impact of multi-epoch training (repeated data). This research demonstrates that while additional compute yields diminishing returns as tokens are repeated, performance can be explicitly modeled by accounting for repetition (Muennighoff et al., 2023). Our work adapts this compute–data perspective to self-supervised vision pretraining. By shifting the focus to redundancy-heavy satellite imagery, we explicitly model the interaction between data **Size (**U**)** and **Density (**R**)**, while introducing **Diversity (**V**)** as a critical third variable to explain the efficiency of the learned representations.

3 Experimental Setup

All pretraining experiments are conducted using the Maxar satellite imagery dataset (MaxarTechnologies, 2022). As discussed in Sec. 1, the dataset is characterized by high redundancy; the same geographic regions are frequently re-imaged over time, which yields substantial near-duplicate content upon tiling. The original images possess resolutions of approximately 17000×17000 pixels or larger. We partition these high-resolution images into non-overlapping crops of 500×500 pixels, resulting in a total of $51{,}197{,}237$ tiles. During pretraining, these tiles are resized to the standard model input resolution of 224×224 pixels, following the iBOT preprocessing pipeline.

Subset Selection: To construct filtered subsets across various scales, we employ a *clustering-based curation* method (Vo et al., 2024). This approach applies hierarchical k-means clustering to representation vectors followed by a balanced sampling strategy to ensure that no single cluster disproportionately dominates the selected subset. Representation vectors for all tiles are computed using the [CLS] embedding from a pretrained DINOv2 model (Oquab et al., 2024). Specifically, we perform clustering across $n = 4$ hierarchical levels with $k \in \{50\text{k}, 10\text{k}, 5\text{k}, 1\text{k}\}$ clusters per level, respectively. By varying the target subset size, we generate multiple curated subsets spanning a broad range of unique-data sizes U. To disentangle the benefits of curation from the effects of subset size alone, we also include a *random subsampling baseline* for several target sizes as a comparative control.

Model Pretraining: We pretrain iBOT with a ViT-B/16 backbone following the standard protocol (Zhou et al., 2021), employing the same multi-resolution augmentations as the original implementation. Optimization is performed using AdamW with a per-GPU batch size of 200 and a world size of 8, resulting in a global batch size of $B = 1600$. All runs are conducted in distributed mode utilizing FlashAttention-2 for computational efficiency.

In this work, we define the pretraining compute budget C as the **total sample exposures** (i.e., the cumulative number of images seen during training, rather than optimizer steps). We conduct pretraining across all subset sizes for $C \in \{10\text{M}, 20\text{M}, 30\text{M}, 40\text{M}, 50\text{M}\}$. For the full dataset and the $U = 750{,}000$ subset, we extend the budget to $C \in \{100\text{M}, 200\text{M}, 300\text{M}, 400\text{M}\}$. Given a global batch size $B = 1600$, the number of optimizer steps S is derived as $S = C/B$ (e.g., 10M exposures $\approx 6{,}250$ steps; 50M exposures $\approx 31{,}250$ steps). For a subset of size U, the *data density* (repetitions) is defined as: $R = \frac{C}{U}$ which corresponds to the number of training epochs. More pretraining detailes are A.1.

We evaluate the quality of the pretrained representations by measuring k-NN classification accuracy across four remote-sensing benchmarks. The details are in Appendix A.2.

4 Proposed Predictive Scaling Law

Each pretraining run produces a tuple (U, C, V, y), where U is the number of unique images in the curated subset, C is the total number of images processed during pretraining (Sec. 3), V is the subset diversity measured by the Vendi score, and y is downstream performance measured as the mean k-NN accuracy averaged over downstream tasks (Sec. 3). Since the backbone and optimization setup are fixed, the dominant degrees of freedom are data and compute allocation.

We quantify subset diversity using the *Vendi score* V (Friedman & Dieng, 2023). For each subset, we uniformly sample m tiles (due to the $O(m^2)$ cost), compute L2-normalized DINOv2 [CLS]

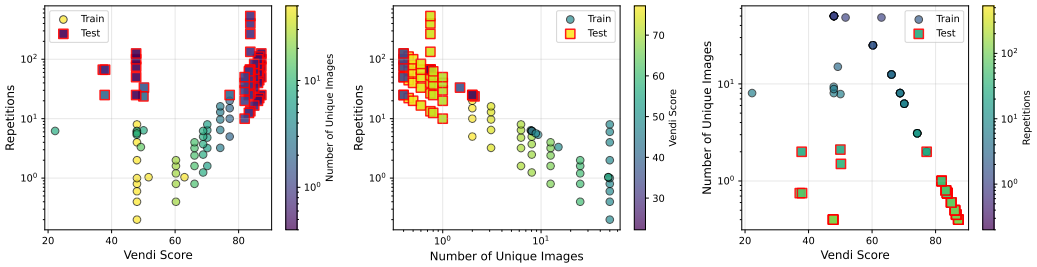

Figure 1: Visualization of the experimental design space and extrapolation split. Each point corresponds to one experiment with Vendi score V, number of unique images U (in millions), and repetition count R (linked through the compute budget $C \approx U \cdot R$). We show pairwise projections of these variables, using the third variable as color: (a) V vs. R (color: U), (b) U vs. R (color: V), and (c) V vs. U (color: R). Markers indicate whether a run belongs to the fit set (Train) or the extrapolation set (Test); log scales are used for U and/or R where indicated.

embeddings, form the cosine-similarity Gram matrix K, normalize $\tilde{K} = K/\mathrm{tr}(K)$, and compute its eigenvalues $\{\lambda_\ell\}$ (with $\sum_\ell \lambda_\ell = 1$). We then define $V = \exp\Big(-\sum_\ell \lambda_\ell \log \lambda_\ell\Big)$, and repeat this computation 10 times with different random seeds, reporting the mean V across subsets.

We fit and compare two functional forms that relate performance to U, R and V.

(L1) Effective-data law. Following (Muennighoff et al., 2023), which showed that training on multiple epochs on the same data brings diminishing returns, we define an effective unique data

$$U_{\text{eff}}(U, R, V) \;=\; U \;+\; U\, x \left(1 - \exp\!\left(-\frac{R\,V}{x}\right)\right), \tag{1}$$

and model performance as

$$\hat{y}_1(U, R, V) \;=\; 1 - E \;-\; \frac{A}{\left(U_{\text{eff}}(U, R, V)\right)^{\beta}}. \tag{2}$$

(L2) Additive-power law. Motivated by compute–data scaling perspectives and power-law behavior in language modeling (Hoffmann et al., 2022), we also fit a simpler additive form with separate contributions from unique data and diversity-weighted density:

$$\hat{y}_2(U, R, V) \;=\; 1 - E \;-\; \frac{A}{(UR)^{\alpha}} \;-\; \frac{B}{(VR)^{\gamma}}. \tag{3}$$

Fitting and evaluation. For each law, we fit parameters by bounded nonlinear least squares (with multiple random restarts), enforcing $E \in [0, 1]$ and positive scales/exponents ($A, B, x, \alpha, \beta, \gamma > 0$). We evaluate predictive accuracy (MSE) and rank correlation on both the fit split and a held-out extrapolation split defined by $R > R_g$ or $V > V_g$ (Sec. 3), and report which law yields more stable fits and better extrapolation. We fit the scaling-law parameters on configurations with at most 20 repetitions and Vendi score up to 80, and evaluate on the held-out configurations that exceed either threshold (see Fig. 1). For a formal paper, you can use the following LaTeX snippet to describe your methodology concisely:

Scaling law models are fitted using the Trust Region Reflective algorithm via `scipy.optimize.curve_fit`, employing bounded nonlinear least-squares optimization. The procedure utilizes model-specific initial estimates and parameter bounds defined in the model registry, with a maximum of 10,000 function evaluations to ensure convergence.

5 Results and Conclusion

In Fig. 3 we show that high quality data subset matters and with the same compute we can get better results with higher diversity subsets. This somewhat contradicts with Goyal et al. (2024), that long

compute time can bring to higher performance in terms of infinite compute. Here we see that Maxar training with the full data saturates after 30M views and drops after 200M views. As one can see in Fig. 2 and Table 1, the Additive power law fits the data much better than the Effective data law. Note that both scaling laws fail to fit runs with relatively small 10M budgets, whereas additive power law shows almost perfect results on runs with higher compute budget. Therefore, the correlation between the VR term and U might have an additive nature. We also tried few other formulas that failed to fit, particularly when V and R were in separate terms. Breaking the symmetry of V and R by adding another parameter $V^{\alpha}R^{\beta}$ did not improve the fit.

We note that the scaling laws were fit on relatively irregular sets of experiments, and adding more data points within the interpolation range might improve the fits.

Future work should focus on testing these ideas and laws on datasets from other domains, such as natural imagery, autonomous driving images, scientific imaging (e.g., microscopy and astronomy).

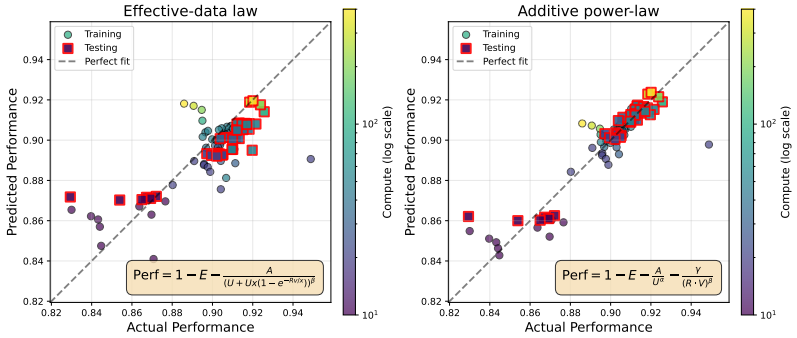

Figure 2: Predicted vs. actual downstream performance for two scaling-law parameterizations. Each point is one training run; circles are runs used to fit the law (Train) and squares are held-out extrapolation runs (Test). The dashed line denotes perfect prediction (y=x). Point color encodes total compute budget (log scale in millions). Left: Effective-data law using an effective unique-data term that saturates with repetitions. Right: Additive power-law baseline combining unique-data and repetition/diversity terms additively.

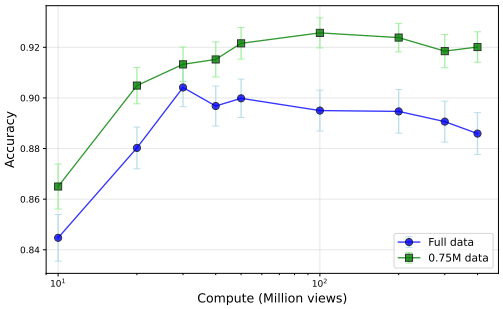

Figure 3: Average k-NN classification performance of iBOT models pretrained on the full Maxar dataset and its best subset (750K).

METRIC	Effective-data	Additive
MSE (train)	0.000227	**0.000108**
Corr (train)	0.7516	**0.8755**
MSE (test)	0.000158	**0.000041**
Corr (test)	0.8233	**0.9482**

Table 1: Comparison of the two proposed scaling law formulations. *Train* refers to the set of the models used for fitting the scaling laws. *Test* indicates extrapolation results.

REFERENCES

Amro Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *CoRR*, abs/2303.09540, 2023. doi: 10.48550/ARXIV.2303.09540. URL https://doi.org/10.48550/arXiv.2303.09540.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson,

Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien 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. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998. URL https://doi.org/10.1109/JPROC.2017.2675998.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/ARXIV.2204.02311. URL https://doi.org/10.48550/arXiv.2204.02311.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. 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*, pp. 22702–22711. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02142. URL https://doi.org/10.1109/CVPR52733.2024.02142.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 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*, pp. 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Jihyeon Lee, Nina R Brooks, Fahim Tajwar, Marshall Burke, Stefano Ermon, David B Lobell, Debashish Biswas, and Stephen P Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17):e2018863118, 2021.

MaxarTechnologies. Open data program. https://registry.opendata.aws/maxar-open-data, 2022. Accessed: 2022 Apr 1.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling data-constrained language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey 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. URL http://papers.nips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Amey P Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3808–3816. PMLR, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/fa0509f4dab6807e2cb465715bf2d249-Abstract-Conference.html.

Hugues Van Assel and Randall Balestriero. A graph matching approach to balanced data subsampling for self-supervised learning. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*.

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *CoRR*, abs/2405.15613, 2024. doi: 10.48550/ARXIV.2405.15613. URL https://doi.org/10.48550/arXiv.2405.15613.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 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)*, pp. 19175–19186, June 2023.

Yi Yang and Shawn D. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In Divyakant Agrawal, Pusheng Zhang, Amr El Abbadi, and Mohamed 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*, pp. 270–279. ACM, 2010. doi: 10.1145/1869790.1869829. URL https://doi.org/10.1145/1869790.1869829.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. ibot: Image BERT pre-training with online tokenizer. *CoRR*, abs/2111.07832, 2021. URL https://arxiv.org/abs/2111.07832.

A Appendix

A.1 Pretraining Detailes

Each sample processing requires approximately 241 GFLOPs, calculated per forward and backward pass for a single 224×224 input.

We adopt the Warmup-Stable-Decay (WSD) scheduler (**?**) for the learning rate, weight decay, and teacher momentum. Specifically, we utilize a linear warmup for the first 625 iterations, followed by a stable phase, and a final linear decay over the terminal 10% of the training budget.

A.2 Model Evaluation

We evaluate the quality of the pretrained representations by measuring k-NN classification accuracy across four remote-sensing benchmarks: UC Merced Land Use (Yang & Newsam, 2010), NWPU-RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), and Brick Kiln (Lee et al., 2021). For each downstream dataset, we extract [CLS] embeddings from the final transformer layer and utilize cosine similarity as the nearest-neighbor metric. To ensure consistency, we determine the optimal value of k for each dataset independently using the model pretrained on the full Maxar dataset; these dataset-specific k values are then held fixed across all other pretraining runs to avoid per-run hyperparameter tuning. Uncertainty is estimated using bootstrap resampling over the test sets with $1,000$ iterations, yielding a mean accuracy m_i and standard deviation σ_i for each dataset $i \in \{1, 2, 3, 4\}$. Our primary performance metric is the unweighted average of the four dataset means:

$$m = \frac{1}{4}\sum_{i=1}^{4} m_i, \quad \sigma = \frac{1}{4}\sqrt{\sum_{i=1}^{4} \sigma_i^2}$$