Less is More? Data Specialization for Self-Supervised Remote Sensing Models

Alvard Barseghyan [*1 2] **Ani Vanyan** [*1 2] **Hakob Tamazyan** [1 2] **Evan Shelhamer** [3] **Hrant Khachatrian** [1 2]

Abstract

Recent foundation models for natural images, such as DINOv2, emphasize data curation as a critical component of the pretraining pipeline. These approaches typically aim to remove near-duplicate images and address semantic imbalance by applying clustering techniques to image representations extracted from pretrained models. While prior work on *data curation* primarily focuses on reducing computational cost while maintaining model quality, in this study we investigate *data specialization*—that is, whether reducing dataset size can improve model quality under a compute-controlled setting. We experiment with two remote sensing datasets, Million-AID and Maxar, apply two data pruning techniques to obtain smaller subsets, and pretrain self-supervision iBOT models while keeping the compute budget constant. We evaluate our models by k-NN on three remote sensing tasks. We show that filtering by hierarchical clustering improves the transfer of Maxar pretraining by 3 percentage points while removing 98.5% of the dataset. On the contrary, neither of the filtering methods improve the transfer of Million-AID pretraining. This motivates future work on identifying and removing "distracting" inputs from the pretraining datasets to improve downstream performance.

1. Introduction

The emergence of foundation models in computer vision has been 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), where large-scale pretraining on raw Internet text data 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 image classification, and pixel-level tasks, such as segmentation.

Recent works, like DINOv2 (Oquab et al., 2024; Vo et al., 2024), highlight the importance of data curation in natural imagery, addressing challenges of data duplication and class imbalance, etc. These approaches leverage feature embeddings extracted by self-supervised networks and employ clustering-based filtering techniques (e.g., k-NN or k-Means) to refine training datasets.

However, most work on **data curation** uses data filtering to enhance the computational efficiency of the pretraining process. The motivation behind the current work is to use filtering methods to **improve model quality**, while using the same amount of compute. We use the term **data specialization** to contrast our goal with that of data curation.

We focus on remote sensing, as there are several publicly available datasets of varying diversity and quality. The goal is to identify "distracting" images in the dataset, that when removed, the downstream performance of the self-supervised models can be improved. Note that the definition of the "distractiveness" of the image is not independent from the dataset; in fact its role might strongly depend on the availability of similar images in the dataset.

We investigate SemDeDup (Abbas et al., 2023) and Hierarchical Clustering (Vo et al., 2024) for the purpose of dataset specialization. Specifically, we pretrain a self-supervised model using the iBOT algorithm (Zhou et al., 2021) on various subsets of the Million-AID and Maxar datasets, deduplicated, and balanced datasets, where we use DINOv2 features for clustering. The impact of these different data versions is evaluated through k-NN classification on three remote sensing classification benchmarks: UC Merced (Yang & Newsam, 2010), RESISC-45 (Cheng et al., 2017), and EuroSAT (Helber et al., 2019).

Furthermore, we synthetically duplicate the filtered datasets for controlled experiments on pretraining. We define data

[*]Equal contribution [1]YerevaNN [2]Yerevan State University [3]The University of British Columbia. Correspondence to: Alvard Barseghyan <alla@yerevann.com>, Ani Vanyan <ani@yerevann.com>.

Proceedings of TerraBytes: Towards global datasets and models for Earth Observation Workshop at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

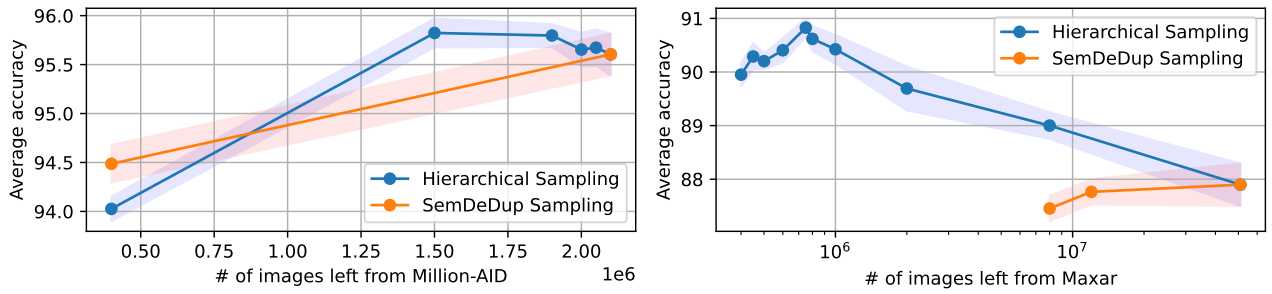


Figure 1. Average k-NN classification performance of iBOT models pretrained on filtered subsets of the Million-AID and Maxar datasets. The model trained on a hierarchically sampled subset of Maxar achieves significantly better performance than the one trained on the full dataset, despite using only 1.5% of the images. All experiments within each plot use the same compute budget.

deduplication as the process of identifying and removing semantically similar or identical data points, **duplication** as the introduction of additional copies of existing data points and **balancing** as the process of adjusting the dataset to promote uniform class distributions (even if the classes are unknown).

Our main finding is that the impact of data specialization techniques varies significantly across datasets. On Million-AID, we observed no substantial improvement compared to using the full dataset. In contrast, on Maxar, subsampling via hierarchical clustering consistently enhanced model quality—even when 98.5% of the images are removed. We hypothesize that the largest clusters in the pretraining dataset, as identified by the hierarchical clustering algorithm, contain many redundant or low-value images that act as distractors, ultimately having a net negative effect on the downstream performance of self-supervised models.

2. Related Work

2.1. Self-Supervised Pretraining

After the promising achievements of self-supervised pretraining on large-scale web-collected data in natural language processing, similar approaches (He et al., 2022; Zhou et al., 2021; Oquab et al., 2024) have gained popularity for natural images. One such method, DINOv2 (Oquab et al., 2024), demonstrates strong performance on downstream tasks, even with a frozen encoder or k-NN classification. For aerial imagery, especially remote sensing images, leveraging self-supervised pretraining is crucial since annotating such data is complex and labor-intensive. As mentioned in some works (Oquab et al., 2024; Abbas et al., 2023; Vo et al., 2024; Dubey et al., 2024), the quality of pretraining data plays a significant role in model performance. These methods apply data curation steps to large-scale raw datasets to balance class distributions and remove duplicates. In contrast, Goyal et al. (2024) argues that when sufficient training compute is available, data typically removed by filtering can actually be beneficial in later training iterations. In this work, we analyze the effect of data curation for pretraining self-supervised models on remote sensing images.

2.2. Data Curation

According to the methods mentioned above, data quality is important for the final model performance. To achieve this, some remote sensing methods, such as SatlasPretrain (Bastani et al., 2023) and GFM (Mendieta et al., 2023), collect data in a curated manner. However, they do not address issues related to duplicates and class imbalance. Identifying duplicates or organizing an unlabeled dataset into meaningful categories is challenging. In DINOv2, this problem is addressed by using another network, SSCD (Pizzi et al., 2022), as a feature extractor and filtering the data based on the relationships between feature vectors. SemDeDup (Abbas et al., 2023), on the other hand, utilizes features provided by the CLIP (Radford et al., 2021) encoder. Hierarchical clustering method (Vo et al., 2024) applies a hierarchical k-means algorithm to sample a desired amount of data from a large-scale dataset while maintaining a balance between clusters. As an alternative to clustering-based sampling, Van Assel & Balestriero reformulated the initial problem as a graph matching task, where the goal is to identify a data subset that is most distinct in terms of pairwise similarities.

3. Data Specialization Experiments

We start from the original versions of Million-AID (Long et al., 2021) and Maxar datasets (MaxarTechnologies, 2022).

Every image in the original dataset was divided into smaller tiles, which gives 2 106 700 images for Million-AID and 51 197 237 for Maxar. To create the Maxar dataset, we used Open Data Program, where images of crisis events can be found in different dates.

3.1. Filtering methods

SemDeDup. SemDeDup is a universal data filtering algorithm designed to remove semantically similar datapoints from large-scale datasets. It can be applied to both images and text. The filtering algorithm follows these steps: **(1)** Extract feature vectors for all datapoints using a feature extractor. We use pretrained DINOv2 (Oquab et al., 2024) of size ViT-B. **(2)** Cluster these feature vectors using the k-means algorithm. **(3)** Compute pairwise cosine similarity within each cluster, considering pairs with a similarity above a defined threshold as duplicates. **(4)** For each group of semantic duplicates, retain the image with the lowest cosine similarity to the cluster centroid.

Note that the clustering step is necessary to reduce computational complexity; however, if sufficient memory is available, it is possible to skip this step and compute pairwise cosine similarity across all datapoints directly. We set $k = 50K$ and use a similarity threshold of 0.8.

To obtain further deduplication, we apply this algorithm N times, iteratively filtering the remaining dataset. For Million-AID we have $414K$ images after $N = 50$ iterations. For Maxar, we have $12M$ images with $N = 1$ iteration, and $8M$ images after $N = 190$ iterations.

Clustering Based Sampling. This algorithm enables sampling a balanced dataset from a large-scale data collection. It applies hierarchical k-means clustering to the given dataset and then samples the requested amount of data in a balanced manner, ensuring that no large semantic categories remain. For our clustering levels, we set $n = 4$ and assign these numbers of clusters at each level: $k = 50K, 10K, 5K, 1K$. The algorithm supports filtering down to a specified number of images. We use different thresholds to obtain multiple subsets of the datasets.

3.2. Model Pretraining and Evaluation

We pretrain our model for a **constant number** of iterations. All models process exactly 50 million images during the pretraining. This implies that the larger subsets are trained for fewer epochs. We follow the original iBOT framework, except for the schedulers. Specifically, we adopt the Warmup-Stable-Decay (WSD) scheduler (Hu et al., 2024) for learning rate, weight decay, and momentum scheduling. Each iteration requires approximately 241 GFLOPs.

We evaluate the pretrained models using k-NN algorithm ($k = 1$) on three classification datasets. We believe k-NN offers a more precise assessment of model quality as it minimizes the confounding factors typically introduced by linear probing or heavier fine-tuning of the models.

We perform test set bootstrapping to obtain mean m_i and standard deviation σ_i of accuracies on all datasets ($i = 1, 2, 3$). For each pretrained model, we report the average of the three means as our main metric, and the normalized square root of the sum of variances on individual datasets as the standard devation: $\sigma = \frac{1}{3}\sqrt{\sum_{i=1}^{3} \sigma_i}$.

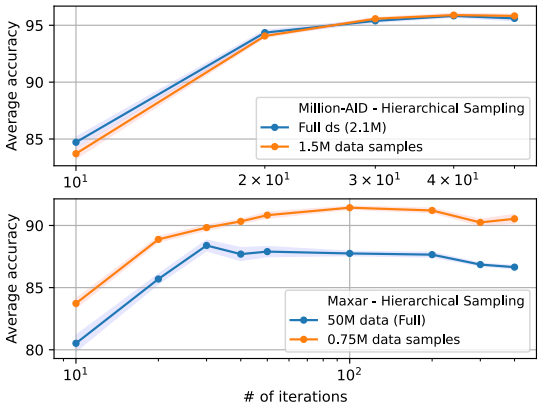


Figure 2. Average k-NN classification performance of iBOT models pretrained on full and the best performed subset of the Million-AID and Maxar datasets.

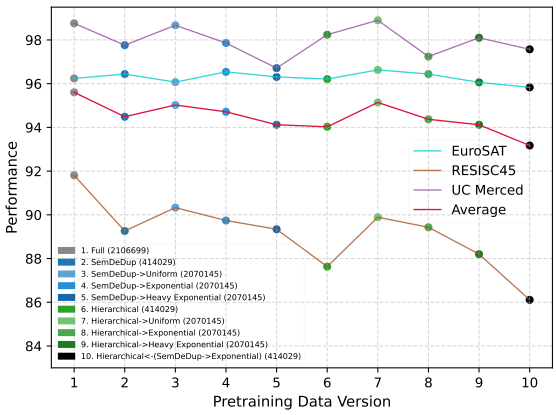
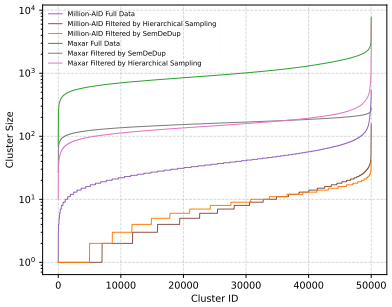


Figure 3. Analysis of the artificially duplicated versions of the filtered Million-AID dataset.

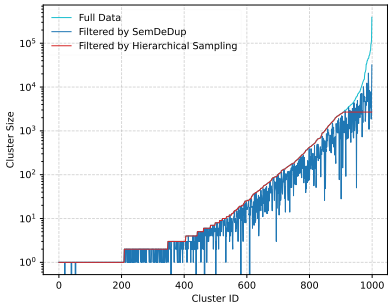
3.3. Results

Figure 1 shows that we were not able to improve downstream performance using SemDeDup filtering for both datasets. Instead, the filtering based on hierarchical clustering produced surprising results. On Million-AID, keeping 1.5M+ images gave a slight improvement in performance, while on Maxar we saw downstream performance improving until up to the removal of 98.5% of the original dataset. The peak performance was observed at 750K images.

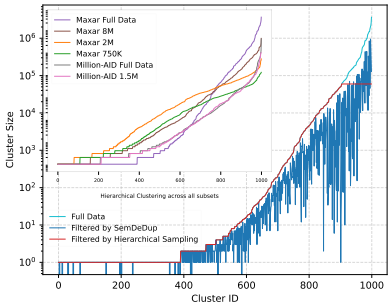
This implies that the **vast majority of the images in Maxar dataset are "distracting"**, i.e. have net negative effect on the pretraining. In Million-AID, the percentage of such im-



(a) Cluster sizes of Million-AID and its filtered versions using SemDeDup's clustering method.



(b) The number of images in each hierarchical cluster obtained from the full Million-AID, for the original dataset and its subsets.



(c) The number of images in each hierarchical cluster obtained from the full Maxar, for the original dataset and its subsets.

Figure 4. Comparison of the filtering methods.

ages is less than 25%, and the performance gain is minimal.

Note that the downstream performance of the models pretrained on any subset of Maxar is significantly lower compared to Million-AID-pretrained models. We believe it is because Million-AID contains significantly wider coverage of possible remote sensing images, but rigorous quantification of such coverage is left for future work.

4. Analysis

These results raise a critical question: is it possible to identify distracting images in large datasets without extensive pretraining experiments? We leave this as an open question, and provide the following visualization that might shed some light on it. We independetly perform hierarchical clustering on the original Million-AID and Maxar datasets, and a few hieararchically filtered versions of the latter. Inner illustration in Fig. 4c shows the cluster sizes for each of the subsets. It is clear that the filtered versions of Maxar have relatively more uniform distribution of cluster sizes. On the other hand, it is hard to see what is the fundamental difference in cluster size distributions of Maxar-2M and Maxar-750K, when the former has many distracting images, while the latter has none we could identify.

We hypothesize that most of the images in the large clusters (as discovered by the hierarchical clustering algorithm) are distractors. To test this hypothesis, we artificially enlarged some of the clusters by controlled duplication of images.

4.1. Controlled Duplication of the Filtered Datasets

We took the well filtered subsets of Million-AID of size 414K, and synthetically enlarged them in uniform and non-uniform manners. First, we generate the $\tilde{5}$ augmentations for each of the images, and call it "uniform". Then, we produce two more datasets of the same size by applying augmentations in a non uniform way: the number of augmentations per image grows exponentially. The two versions, named "exponential" and "heavy exponential" differ by the coefficents of the exponential function that determines the number of augmentations of each image.The augmentation functions are chosen from a set of ten weak augmentations, such as flipping and rotation. Then we pretrain iBOT on each of these datasets and evaluate as in the previous sections.

Figure 3 shows that uniformly duplicated versions of the filtering datasets perform better than the pure filtered ones. This can be explained by the additional information introduced by image augmentations. Once the duplication is performed non-uniformly, the positive effects from the augmentations are outweighted by the negative effects of the distrating images. The heavier is the non-uniformity, the lower is the downstream performance of the models. These experiments provide some support for our hypothesis on the negative impact of too large clusters in the pretraining datasets.

4.2. Comparison of the Filtering Methods

Finally, we explore whether the two filtering methods perform similarly. First, we take the original Million-AID dataset, cluster it as in the first step of SemDeDup, and plot the sizes of each of the 50K clusters (Fig. 4a). Then we obtain filtered subsets with 414K images from Million-AID using both algorithms, independently apply the same clustering algorithm on the new smaller datasets and plot the cluster sizes. We can see that even after $N = 50$ iterations of SemDeDup, there are still a few relatively large clusters. We see that there are few but much larger clusters in the subset obtained using hierarchical clustering.

Fig. 4b and 4c show the number of remaining images in the original clusters after both methods of filtering. Hierarchical method enforces all clusters to have fewer images than a

fixed threshold. The subset filtered by SemDeDup does not follow the threshold for most of the (original) large clusters in both datasets.

Conclusion

In this work, we show that self-supervised pruning can improve representation learning in remote sensing under a fixed compute by removing "distracting" samples. Hierarchical clustering boosts the downstream performance on both datasets, though the magnitude of the improvement depends on the diversity of the initial dataset. These findings highlight the potential of dataset-specific filtering, and motivate future work on automatic identification of harmful pretraining samples.

Acknowledgements

The research was supported by the Higher Education and Science Committee of MESCS RA (Research project No 24RL-1B049). This work was supported by the Strategic Armenian Science and Technology Investment Community (SASTIC).

References

Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. 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.

Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., and Kembhavi, A. 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*, pp. 16726–16736. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01538. URL https://doi.org/10.1109/ICCV51070.2023.01538.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (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.

Cheng, G., Han, J., and Lu, X. 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.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. 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.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Di-

nan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Goyal, S., Maini, P., Lipton, Z. C., Raghunathan, A., and Kolter, J. Z. 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.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. 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.

Helber, P., Bischke, B., Dengel, A., and Borth, D. 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.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. 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.

Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., Zhang, X., Thai, Z. L., Zhang, K., Wang, C., Yao, Y., Zhao, C., Zhou, J., Cai, J., Zhai, Z., Ding, N., Jia, C., Zeng, G., Li, D., Liu, Z., and Sun, M. Minicpm: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395, 2024. doi: 10.48550/ARXIV.2404.06395. URL https://doi.org/10.48550/arXiv.2404.06395.

Long, Y., Xia, G., Li, S., Yang, W., Yang, M. Y., Zhu, X. X., Zhang, L., and Li, D. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 14:4205–4230, 2021. doi: 10.1109/JSTARS.2021.3070368. URL https://doi.org/10.1109/JSTARS.2021.3070368.

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

Mendieta, M., Han, B., Shi, X., Zhu, Y., and Chen, C. Towards geospatial foundation models via continual pretraining. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 16760–16770. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01541. URL https://doi.org/10.1109/ICCV51070.2023.01541.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., and Douze, M. A self-supervised descriptor for image copy detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14512–14522. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01413. URL https://doi.org/10.1109/CVPR52688.2022.01413.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 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.

Van Assel, H. and Balestriero, R. A graph matching approach to balanced data sub-sampling for self-supervised learning. In *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*.

Vo, H. V., Khalidov, V., Darcet, T., Moutakanni, T., Smetanin, N., Szafraniec, M., Touvron, H., Couprie, C., Oquab, M., Joulin, A., Jégou, H., Labatut, P., and Bojanowski, P. 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.

Yang, Y. and Newsam, S. D. Bag-of-visual-words and spatial extensions for land-use classification. In Agrawal, D., Zhang, P., Abbadi, A. E., and Mokbel, M. F. (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.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A. L., and Kong, T. ibot: Image BERT pre-training with online tokenizer. *CoRR*, abs/2111.07832, 2021. URL https://arxiv.org/abs/2111.07832.