THE ROLE OF LANGUAGE IMAGE PRE-TRAINING DATA IN TRANSFER LEARNING

Rahim Entezari[1], **Mitchell Wortsman**[2], **Olga Saukh**[1], **M. Moein Shariatnia**[4] **& Hanie Seghi**[3], **& Ludwig Schmidt**[2]
TU Graz/CSH Vienna[1] University of Washington[2] Google Brain[3] Tehran University of Medical Sciences[4]

ABSTRACT

We explore which pre-training dataset should be used to achieve the best transfer learning performance. We investigate the impact of pre-training on the few-shot and full fine-tuning performance using 7 pre-training datasets, and 9 downstream datasets. Through extensive controlled experiments, we find that the choice of the pre-training dataset is essential for the few-shot transfer, but its role decreases as more data is made available for fine-tuning. Additionally, we explore the role of data curation and examine the trade-offs between label noise and the size of the pre-training dataset. We find that using 2000× more pre-training data from LAION can match the performance of supervised ImageNet pre-training. [1]

1 INTRODUCTION

The best-performing computer vision models are produced by the transfer learning paradigm. While transfer learning is not new, the substantial improvement in the quality of the pre-trained models in recent years has brought transfer learning to the spotlight (e.g., CLIP (1), BASIC (2), and Flamingo (3)). These improvements are driven by new datasets for pre-training as well as better pre-training algorithms. This naturally leads to a question:

How do the dataset and the algorithm used for pre-training affect downstream performance?

In contrast to prior works (4; 5; 6; 7; 8; 9), our main focus is on the role of the pre-training data distribution in downstream performance. We set up systematic experiments to explore our research questions and contributions as follows:

Do different pre-training distributions lead to different transfer learning performances? In practice, one has many options to download pre-training checkpoints and fine-tune the model on the target dataset. Should we expect different pre-training datasets to perform differently in the transfer setting? When controlling for the size of the pre-train model and the downstream dataset, but changing the pre-train dataset, we observe noticeable differences in downstream accuracy in the few-shot setting (only a few examples per class are available for fine-tuning). However as more samples are available for fine-tuning, the difference in absolute accuracy when varying the pre-training dataset largely evaporates. In the few-shot regime, we observe that certain pre-training datasets (*e.g.*, Shutterstock) consistently lead to a better transfer accuracy than the other (*e.g.*, WiT) across many downstream tasks. However, the ranking of the other pre-training datasets in our selection appears mixed. Moreover, even the pre-training dataset which leads to the worst transfer accuracy (WiT) still outperforms training from scratch (see Section 3.1, Figure 1 and Figure 6).

How much is expensive labeling worth compared to noisier but larger pre-training data? We compare different pre-training strategies: supervised pre-training on the carefully labeled ImageNet dataset and semi-supervised pre-training on language-image pairs from larger but noisier datasets. We find that pre-training on a well-curated dataset leads to better transfer accuracy than pre-training on a noisy dataset of a similar size. Our investigations also show that pre-training on a 15x-2000x larger but noisier dataset (LAION) can close the gap for some downstream tasks (see section 3.4, section 3.5, Figure 2 and Figure 3).

[1] The code is available here `https://github.com/rahimentezari/DataDistributionTransferLearning`

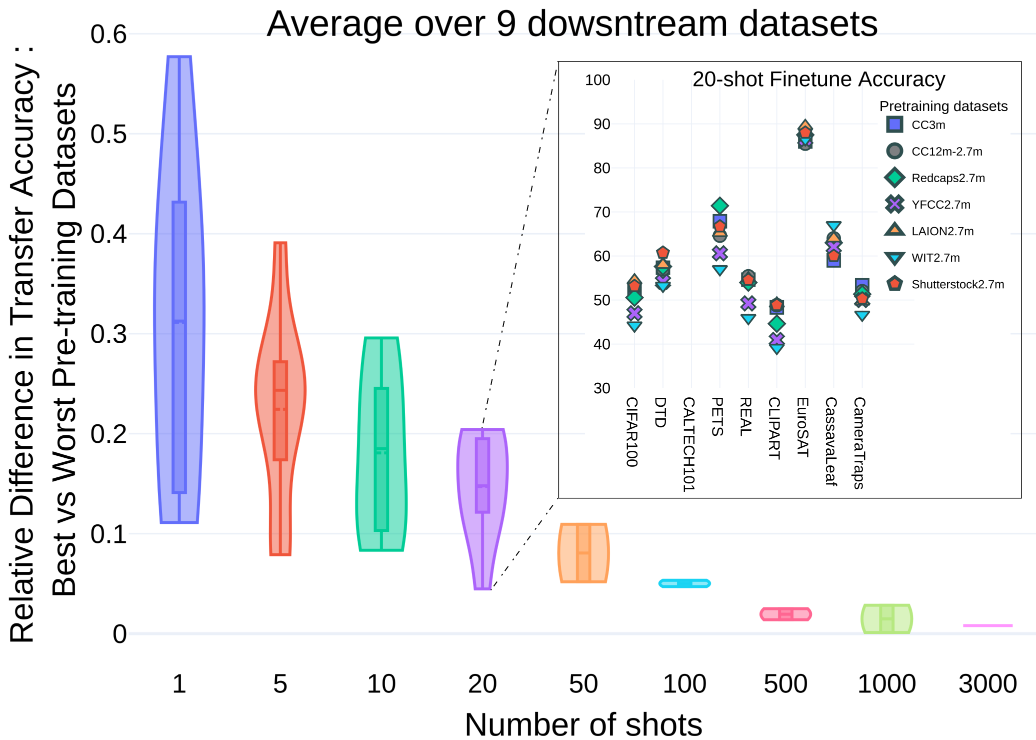


Figure 1: **Differences between various pre-training sources diminish as more data is available for the downstream tasks.** In the few-shot setting, different pre-training datasets lead to noticeable differences in downstream performance. However, if many samples are available for fine-tuning, the difference in absolute accuracy between models pre-trained on different sources largely evaporates (see Figure 6 for a detailed view).

We conduct an extensive empirical investigation in the context of transfer learning for computer vision tasks (See section D for details on 4000 experiments). Our study covers seven pre-training datasets including YFCC, LAION, Redcaps, Conceptual Captions-3m, Conceptual Captions-12m, WiT, Shutterstock, and ImageNet (10; 11; 12; 13; 14; 15; 16), nine fine-tuning datasets including CIFAR100, DTD, Caltech-101, Oxford-PETS, REAL and CLIPART from DomainNet, EuroSAT, Cassava Leaf Disease, and Caltech Camera Traps (17; 18; 19; 20; 21; 22; 23; 24) with CLIP pre-training (1). To evaluate downstream performance, we examine both few-shot fine-tuning and full fine-tuning.

We review closely related works in Appendix Section A. Section 2 explains our experimental setup. Section 3 details our observations relating to our research questions. Due to space limits, we discuss our findings and conclude with future research directions in Appendix Section B.

2 Experimental Setup

Model The main focus of this study is the CLIP model (1). This model has demonstrated unprecedented robustness to natural distribution shifts (25; 26), and transfers well to many downstream tasks (1; 27). Given an image-text pair, CLIP learns a joint embedding space for both images and their captions and tries to maximize the cosine similarity between the text and image embedding for an image relative to the cosine similarity of unaligned pairs. We use the CLIP implementation from the OpenCLIP GitHub repository (28).

Pre-training and Fine-tuning We mainly use ResNet-50 (29) as the image encoder unless stated otherwise. We vary the pre-training data distribution in section 3.1, curation method in section 3.4.

For most of the experiments, we fine-tune the pre-trained model end-to-end on the target transfer dataset unless stated otherwise. For each pre-trained model and downstream transfer dataset, we used a large grid search over various fine-tuning hyperparameters including learning rate, batch

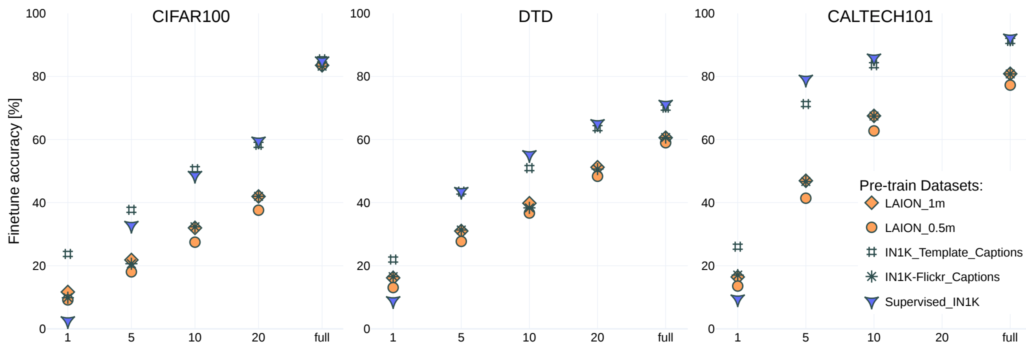


Figure 2: **Effect of data curation and labeling.** Supervised pre-training on ImageNet leads to better transfer accuracy than self-supervised pre-training (IN1K-Template-Captions). On a different comparison between ImageNet and LAION distributions, pre-training CLIP on ImageNet (with clean template captions) outperforms LAION-1m by a large margin. See Figure 4 for other datasets.

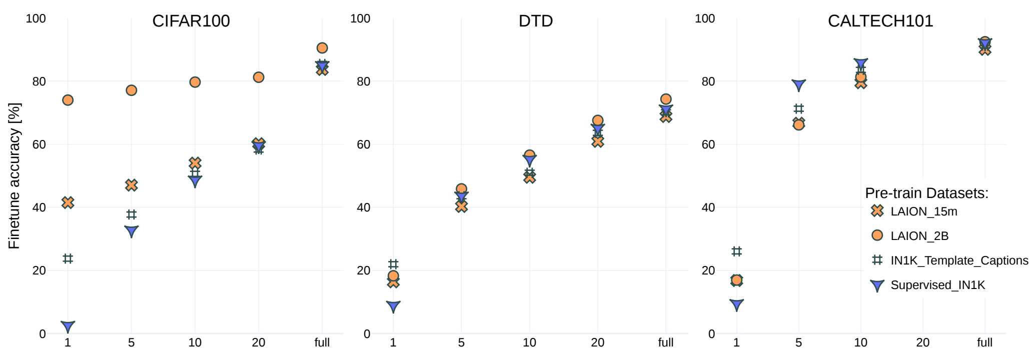


Figure 3: **How much LAION data is worth of ImageNet pre-training?** Including 15x more data from LAION outperform ImageNet pre-training with template captions on CIFAR100. However, DTD, REAL, and CLIPART need 2000x more data from LAION to match or outperform ImageNet pre-training. Even including 2000x more data did not help CALTECH101 and PETS.See Figure 5 for other datasets.

size, and the number of epochs. We report the best-performing accuracy in the plots. Further training details are in Appendix D.

Datasets Our large-scale experiments yield more than 4000 trained networks. Our pre-training datasets cosists of million-size image and language pairs from multiple recent multi-modal datasets including YFCC, LAION, RedCaps, Shutterstock, Conceptual Captions, WiT (10; 11; 12; 13; 14; 15) For downstream tasks, we use nine different datasets CIFAR100, DTD, Caltech-101, Oxford-PETS, REAL, and CLIPART from DomainNet, EuroSAT, Cassava Leaf Disease, and Caltech Camera Traps (17; 18; 19; 20; 21; 22; 23; 24). See Appendix section I for details on pre-training and downstream datasets.

3 Experiments and Results

3.1 What is the impact of different pre-training data sources on transfer learning?

Do we expect different distributions to perform differently in the transfer setting? Figure 1 aggregates transfer performance from different pre-training datasets across all downstream datasets. To get each point, we (1) pre-train CLIP models using a set of seven large sources, (2) fine-tune each pre-trained model on all downstream datasets across different shots (a sweep over multiple hyperparameters, see Appendix D), and (3) for each downstream dataset, calculate the difference between the best and worst fine-tune performance among used pre-training sources, normalized by the maximum fine-tune performance. Figure 1 aggregates over all downstream datasets for each number of shots, highlighting as an example different pre-training models fine-tuned using 20 samples/class on all downstream datasets. We observe that changing the pre-training dataset leads to noticeable differences in the downstream performance in a few-shot setting. However, as more images are available for fine-tuning, the difference in absolute accuracy between different pre-training models is largely diminished. Figure 6 in Appendix shows this diminishing effect in detail for different downstream datasets. The full fine-tuned models have very similar downstream performances despite different pre-training datasets (see the top-right point of CIFAR100 and REAL in Figure 6, and also the top-right point for CameraTraps, Cassava Leaf, and EuroSAT in Figure 6). However, this is not true for DTD, CALTECH101, PETS, and CLIPART, where they have far fewer images per class for fine-tuning on the full dataset. Appendix F extends our results to Vision Transformers (30) instead of ResNet-50.

3.2 Which data distribution is better for transfer learning?

The results presented in Figure 6 demonstrate that pre-training on the Shutterstock and LAION datasets results in superior transfer performance across a range of downstream tasks. A closer look shows the superior performance of Redcaps for PETS. We investigate this further and inspect many pets by looking at random samples from Redcaps at Figure 10. We also look into the most common words in the captions of these pre-training datasets, summarized in Table 3. We observe that ”cats” and ”dogs” are among the most common words in the Redcaps dataset. Table 3 also shows that ”background”, ”design”, ”pattern”, and ”texture” are among the most common words in the captions of Shutterstock, supporting a high correlation to DTD (Describable Textures Dataset). WiT yields the worst performance in most cases because both captions and images (Figure 14) are related to topics about people and geography that are far from the studied downstream tasks.

3.3 How much pre-training contributes to downstream performance as opposed to training from scratch?

While transfer learning from a large pre-training dataset outperforms training from scratch for all downstream tasks, the magnitude of the improvement varies for different datasets in Figure 6. We observe a large improvement for PETS, CALTECH-101, and CLIPART. PETS for example has a small number of samples per class for training (30), which makes it hard to train from scratch. It is also scraped from the web (Google search (19)), similar to our web-scraped pre-training sources. We also hypothesize that a pre-training shows the best improvement when increases both diversity(how hard pre-train data is to fit) and affinity(how pre-training shifts the decision boundary of the scratch model) (31), meaning it should be semantically close to the classes of target task while enriching the distribution over the samples.

3.4 Do well-curated pre-training datasets lead to better transfer?

There has been a significant effort to create computer vision datasets with high-quality labels. On the other hand, many recent datasets are large but noisy. In this section, we are going to investigate: *How much is laborious ImageNet labeling worth?*

To answer this question, we first start by pre-training ResNet-50 on Large Scale Visual Recognition Challenge (ILSVRC) 2012 (32), known as ImageNet-1K, using supervised cross-entropy loss and fine-tune on our downstream datasets in Figure 2. To investigate the role of supervision, we then discard ImageNet labels and use CLIP to pre-train on ImageNet. Because the ImageNet dataset has no captions, we include original Flickr captions, which reduces the size of the image and captions to 0.5M samples (Appendix E describes the required steps to create ImageNet-Flickr). Figure 2 shows that supervised pre-training on ImageNet outperforms CLIP pre-training on ImageNet with Flickr captions by a large margin in all downstream tasks.

However, such a gap could be attributed to two differences between mentioned pre-trainings: (1) supervised vs. contrastive image-language loss, and (2) the size of training samples for supervised-ImageNet (1.2m) is two times larger than CLIP with ImageNet-Flickr captions (0.5m). To remove the second effect we then use all the images from ImageNet, paired with templated clean captions, e.g., “a photo of a *class name*”. This allows us to have a fair comparison between supervised and CLIP pre-training on ImageNet, given the same size. Figure 2 shows that pre-training with clean captions improves the performance of CLIP pre-training by a large margin and outperforms supervised pre-training on CIFAR100. However, supervised pre-training on ImageNet still performs best for the rest of the other datasets.

3.5 How much LAION data is the ImageNet pre-training worth?

Figure 2 compares the ImageNet distribution with LAION. Pre-training CLIP on the ImageNet distribution (with template captions) outperforms LAION-1m by a large margin. Findings from Figure 1 with the same pre-training loss are now extended to different losses in Figure 2, *i.e.*, the gap between the supervised ImageNet (with template captions) pre-training and the contrastive LAION-1m pre-training shrinks as more data for the downstream task are available. Interestingly, pre-training CLIP on LAION-1m is only as good as ImageNet with Flickr captions with half of the data. We also scale LAION pre-training size in Figure 3 to see if LAION can outperform ImageNet pre-training and downstream performance. Figure 3 shows that including 15× more data from LAION outperforms ImageNet pre-training with template captions only on CIFAR100. However, DTD, REAL, and CLIPART need 2000× more data from LAION to match or outperform ImageNet pre-training. Even including 2000× more data did not help CALTECH101. ImageNet pre-training also outperforms LAION-2B on PETS by a large margin. This is probably because PETS and ImageNet both share many samples of pets like dog breeds.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[2] H. Pham, Z. Dai, G. Ghiasi, H. Liu, A. W. Yu, M.-T. Luong, M. Tan, and Q. V. Le, "Combined scaling for zero-shot transfer learning," *arXiv preprint arXiv:2111.10050*, 2021.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[4] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

[5] S. Abnar, M. Dehghani, B. Neyshabur, and H. Sedghi, "Exploring the limits of large scale pre-training," *arXiv preprint arXiv:2110.02095*, 2021.

[6] K. You, Y. Liu, J. Wang, and M. Long, "Logme: Practical assessment of pre-trained models for transfer learning," in *International Conference on Machine Learning*, pp. 12133–12143, PMLR, 2021.

[7] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau, "Leep: A new measure to evaluate transferability of learned representations," in *International Conference on Machine Learning*, pp. 7294–7305, PMLR, 2020.

[8] A. Deshpande, A. Achille, A. Ravichandran, H. Li, L. Zancato, C. Fowlkes, R. Bhotika, S. Soatto, and P. Perona, "A linearized framework and a new benchmark for model selection for fine-tuning," *arXiv preprint arXiv:2102.00084*, 2021.

[9] D. Bolya, R. Mittapalli, and J. Hoffman, "Scalable diverse model selection for accessible transfer learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19301–19312, 2021.

[10] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[11] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[12] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv preprint arXiv:2111.11431*, 2021.

[13] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

[14] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

[15] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2443–2449, 2021.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[17] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-100 and cifar-10 (canadian institute for advanced research)," 2009. MIT License.

[18] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178, IEEE, 2004.

[20] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505, IEEE, 2012.

[21] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

[22] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[23] "Makerere University AI Lab. Cassava leaf disease classification, 2021." `https://www.kaggle.com/competitions/cassava-leaf-disease-classification/overview`. Accessed: 2022-10-20.

[24] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

[25] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18583–18599, 2020.

[26] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt, "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization," in *International Conference on Machine Learning*, pp. 7721–7735, PMLR, 2021.

[27] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. Gontijo-Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," *arXiv preprint arXiv:2109.01903*, 2021. `https://arxiv.org/abs/2109.01903`.

[28] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," July 2021. If you use this software, please cite it as below.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

[31] R. Gontijo-Lopes, S. Smullin, E. D. Cubuk, and E. Dyer, "Tradeoffs in data augmentation: An empirical study," in *International Conference on Learning Representations*, 2020.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[33] D. Kim, K. Wang, S. Sclaroff, and K. Saenko, "A broad study of pre-training for domain generalization and adaptation," *arXiv preprint arXiv:2203.11819*, 2022.

[34] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[36] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[37] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," *arXiv preprint arXiv:2112.12750*, 2021.

[38] A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt, "Data determines distributional robustness in contrastive language image pre-training (clip)," *arXiv preprint arXiv:2205.01397*, 2022.

[39] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[40] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?," *Advances in neural information processing systems*, vol. 33, pp. 512–523, 2020.

[41] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[42] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021.

[43] A. Islam, C.-F. R. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris, "A broad study on the transferability of visual representations with contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8845–8855, 2021.

[44] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, *et al.*, "Self-supervised pretraining of visual features in the wild," *arXiv preprint arXiv:2103.01988*, 2021.

[45] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, pp. 4904–4916, PMLR, 2021.

[46] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

[47] T. Nguyen, G. Ilharco, M. Wortsman, S. Oh, and L. Schmidt, "Quality not quantity: On the interaction between dataset design and robustness of clip," *arXiv preprint arXiv:2208.05516*, 2022.

[48] S. Santurkar, Y. Dubois, R. Taori, P. Liang, and T. Hashimoto, "Is a caption worth a thousand images? a controlled study for representation learning," *arXiv preprint arXiv:2207.07635*, 2022.

[49] "Common crawl." `https://commoncrawl.org/`. Accessed: 2022-09-20.

APPENDIX

A RELATED WORK

This work is inspired by and closely related to **(author?)** (33) and **(author?)** (5). **(author?)** (33) conducted an in-depth study of the effect of the network architecture, pre-training dataset, supervised vs self-supervised learning objectives, and different domain transfer methods on the transferability of representations to new domains. They found that the transferability of the pre-trained representations depends on factors such as the target benchmark, adaptation method, and network depth. However, they do not study few-shot transfer (where we see the most impact of the pre-training distribution). They also did not provide a set of controlled experiments for some of the studied impacting factors because they are limited to available pre-trained models. For example, when comparing the role of data distribution (their Figure 2, ImageNet-22K vs. JFT-300m), they change the dataset size and also architecture, and the reader is left wondering if JFT has a better distribution for transfer or if the observed effects come from more data or a better architecture?

(author?) (5) also explored how different upstream training settings affect transfer accuracy for two upstream datasets and more than 20 downstream tasks. They showed that as the upstream accuracy increases, the transfer learning performance on downstream datasets saturates. However, the authors study only upstream models that are pre-trained with a supervised loss function on ImageNet-21K (16) or JFT-300M (34) (different size and distributions). In this work, we extend these results to more pre-training datasets and methods, with a special focus on data distribution and curation. **(author?)** (5) also lacks controlled comparison between different distributions in the pre-training datasets *e.g.*, they compare JFT and ImageNet with very different sample sizes. We consider full fine-tuning in addition to few-shot transfer. Moreover, (6; 7; 8) develop metrics for predicting the transferability of a model. Their main focus is to develop a measure to predict the full fine-tune accuracy without actually fine-tuning on the downstream task. While we also cover full fine-tune accuracy, our main research question lies in studying the extent to which pre-training data affect transfer accuracy. Looking at few-shot and full-shot also gives us the ability to study the effect of transfer learning as more target data become available. Moreover, predictability of the transfer performance is mostly limited to supervised ImageNet-1K pretraining, while we scale both pre-training distributions, size, and pre-training loss functions. Transferability line of research also mainly focuses on Internet-crawled datasets, while we extended our results to domain-specific datasets (Camera Traps, Cassava Leaf Diseases, and EuroSAT), Section H extends related works.

B DISCUSSION, LIMITATIONS, AND FUTURE WORK

Discussion As better pre-trained models become available, and more workloads shift from training from scratch to fine-tuning, understanding the transfer learning paradigm becomes increasingly important. Presumably, in the future, a sea of pre-trained models will be available for download from the Internet. Therefore, researchers and practitioners will be faced with the question of where to begin. It will be important to make this choice well, but also to understand to what extent this choice matters. Overall we have observed that different pre-training distributions and methods can lead to differences in downstream transfer accuracy. However, these differences are the largest in the few-shot transfer regime. If many images are used for fine-tuning these differences are mostly diminished. Moreover, while different pre-training decisions lead to similar accuracy in the high-shot regime, they still outperform training from scratch in the setting we consider.

Limitations and Future work There are a number of limitations in our study. For one, we consider only end-to-end fine-tuning, because this method produces the highest accuracy. However, if compute is limited, one may choose to instead use linear probing or other lightweight fine-tuning methods. So far this is not addressed in our study. Another limitation is that we did not do an exhaustive hyperparameter sweep for pre-training. While fine-tuning is cheaper and we are therefore able to do a grid search, for pre-training we are mostly limited to using existing checkpoints. While we think that this reflects a realistic setting, in the future we wish to also better understand the role of hyperparameters.

In addition to the mentioned limitations, future works might include extending experiments to include different samples of ImageNet. One example may include subsets of ImageNet-21K (2.7m in

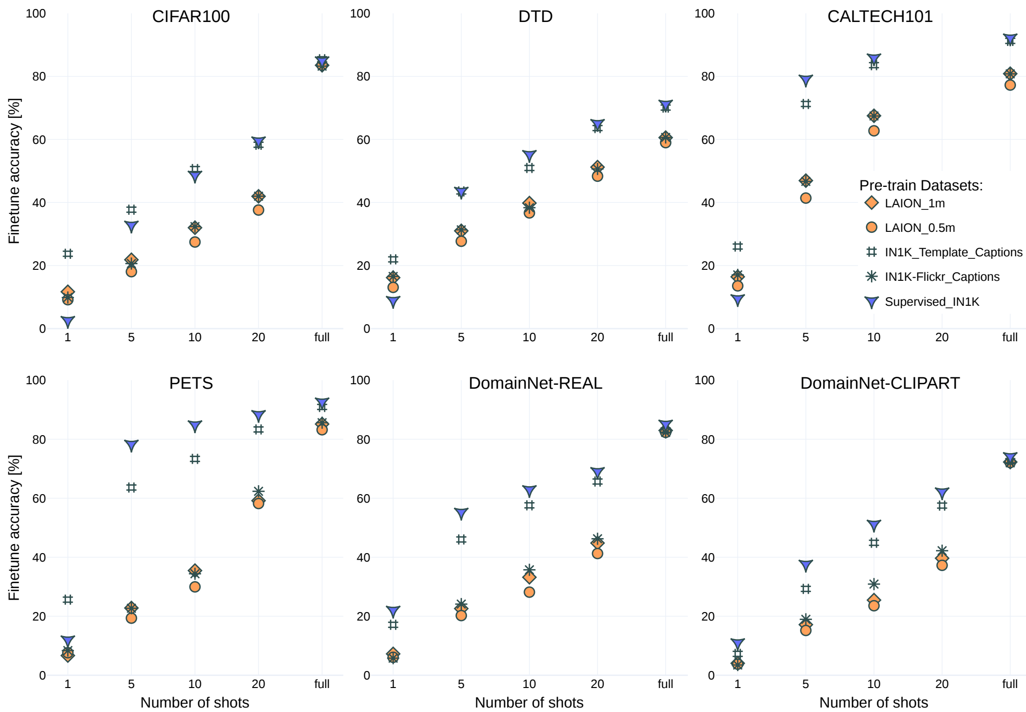


Figure 4: **Effect of data curation and labeling.** Supervised pre-training on ImageNet leads to better transfer accuracy than self-supervised pre-training (IN1K-Template-Captions). On a different comparison between ImageNet and LAION distributions, pre-training CLIP on ImageNet (with clean template captions) outperforms LAION-1m by a large margin.

Figure 6 and 15m in Figure 3) and respective comparison to Shutterstock and LAION distributions. Given our observation of the role of data curation, we also hope that our findings stimulate further direction toward creative methods for dataset curation.

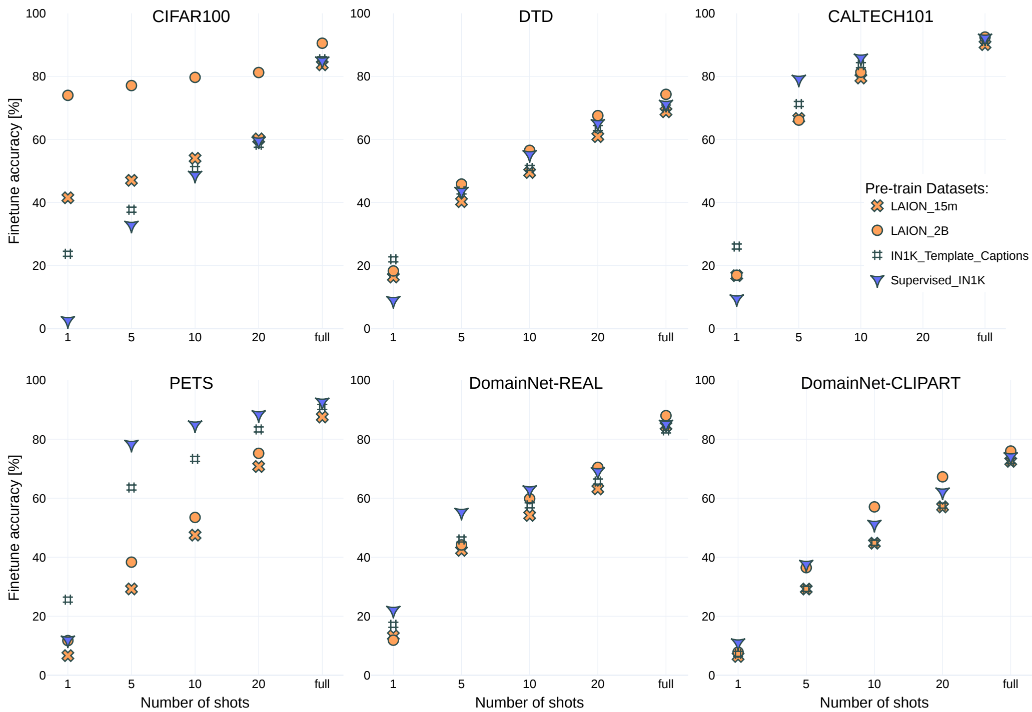
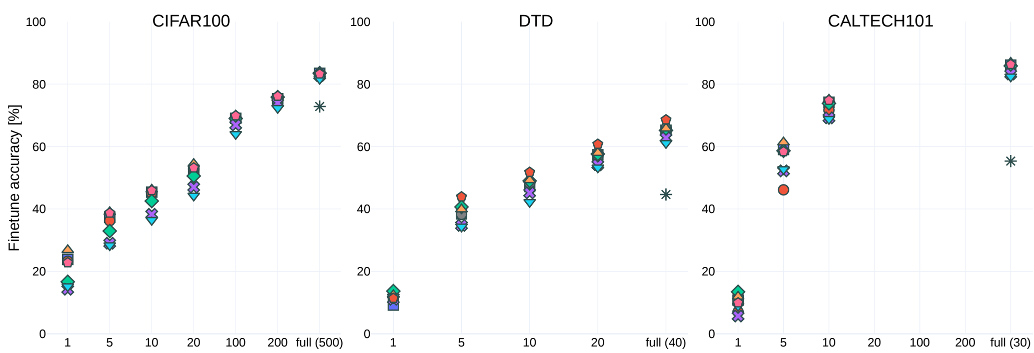
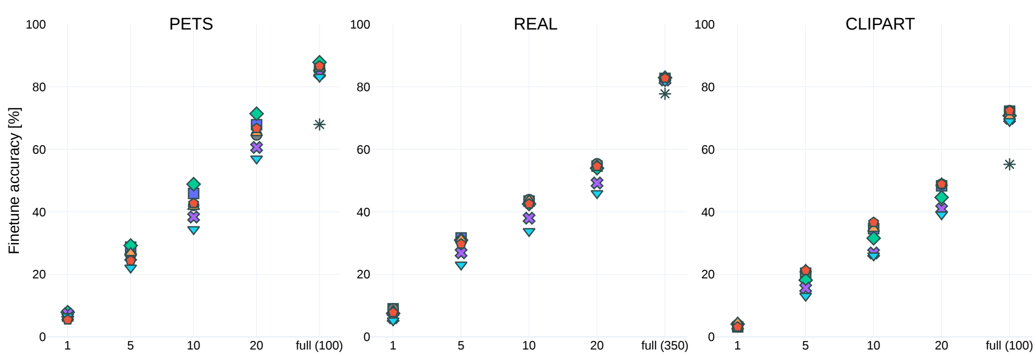


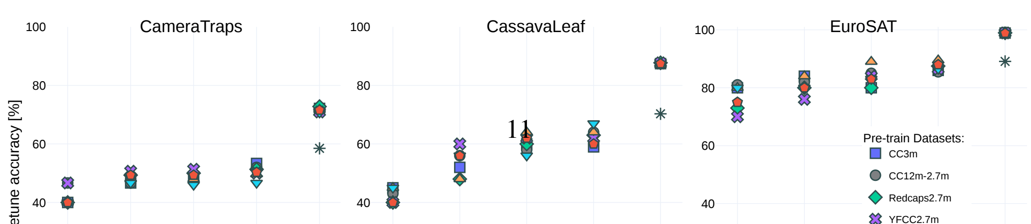
Figure 5: **How much LAION data is worth of ImageNet pre-training?** Including 15x more data from LAION outperform ImageNet pre-training with template captions on CIFAR100. However, DTD, REAL, and CLIPART need 2000x more data from LAION to match or outperform ImageNet pre-training. Even including 2000x more data did not help CALTECH101 and PETS.



(a) CIFAR100 (b) DTD (c) CALTECH101



(d) PETS (e) DomainNet REAL (f) DomainNet CLIPART



C Effect of the pre-training data distribution

Figure 6 shows a detailed for aggregated results shown in Figure 1. In the low-shot setting, different pre-training datasets lead to noticeable differences in downstream performance. If many samples are available for fine-tuning, the difference in absolute accuracy between models pre-trained on different sources largely evaporates.

Figure 6 compares different data sources for pre-training. While Shutterstock shows superior performance on the first six datasets (except for PETS), the best pre-training distribution changes between Camera Traps, Cassava Leaf, and EuroSAT. Changing the pre-training dataset leads to noticeable differences in the downstream low-shot performance of nine datasets.

D Training Details

D.1 CLIP training

Our CLIP models are trained from scratch on each of the pre-training datasets unless otherwise mentioned and follow the training code from the OpenCLIP GitHub repository(28). CLIP models are trained using AdamW optimizer (35) with default PyTorch parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, batch size 1024, and weight decay of 0.1. For learning rate, we start with a learning rate of 10^{-3} and apply a cosine-annealing learning rate schedule (36) with 5,000 steps warm-up. We use the same data augmentations as in(1).

D.2 SimCLR training

Our SimCLR implementation closely follows the training code from the SLIP(37). SimCLR models are also trained for 16 epochs from scratch using AdamW optimizer (35) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$, batch size 1024, and weight decay of 0.1. we start with a learning rate of 10^{-3} and apply a cosine-annealing learning rate schedule (36) with 2 epochs of warm-up. The hidden dimension of SimCLR MLP projection head is set to 4,094 and the output embedding dimension of MLP projection head is set to 256.

D.3 Finetuning details

Each pretrained model is finetuned on the specific downstream task for 128 epochs while the learning rate is mostly from 0.0001, 0.0003, 0.001, 0.003 as starting and applying a cosine-annealing learning rate schedule (36) with 500 steps warm-up and batch size of 128. For each fine-tuning, we choose the best-performing result on the test set among the performed grid search. We use the implementation from the WiSE-FT GitHub repository for fine-tuning, where we have only one model and $\alpha = 1$ (27). For a list of all **4000 experiments**, including their hyperparameters and performance see `https://github.com/AnonymousMLSubmission/DataDistributionTransfer/blob/main/Hyperparameters_results.csv`

E Effect of data curation: ImageNet captioning

We compare CLIP models pre-trained on LAION with CLIP models pre-trained on the following two versions of the curated ImageNet dataset:

- IN1K-Flickr-Captions: This is a subset of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 training set, paired with the original image title, description, and tags from Flickr. Therefore, we can use it for CLIP pre-training. To construct this dataset, **(author?)** (38) start from 14,197,122 image URLs in the ImageNet fall 2011 release, and filter to only include images from Flickr. Next, they restrict the images to the 1,000 classes included in the 2012 ImageNet competition, run the image deduplication routine, and remove text containing profanity. As a result, the dataset of 463,622 images is left along with the newly obtained corresponding text data.

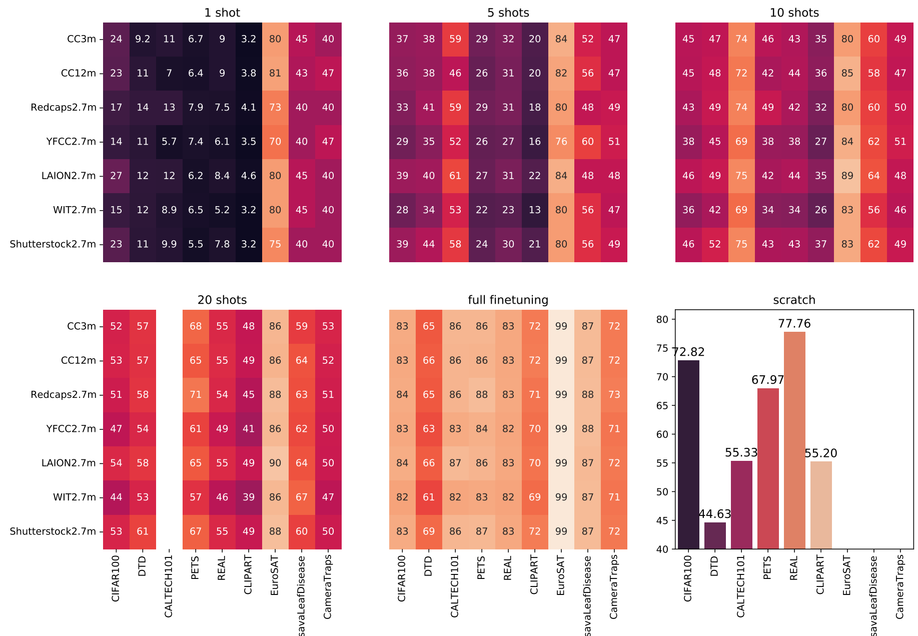


Figure 7: **Effect of pre-training data distribution: a better view.** We change the presentation of Figure 6 for a better view of exact performance numbers on different data distributions and datasets.

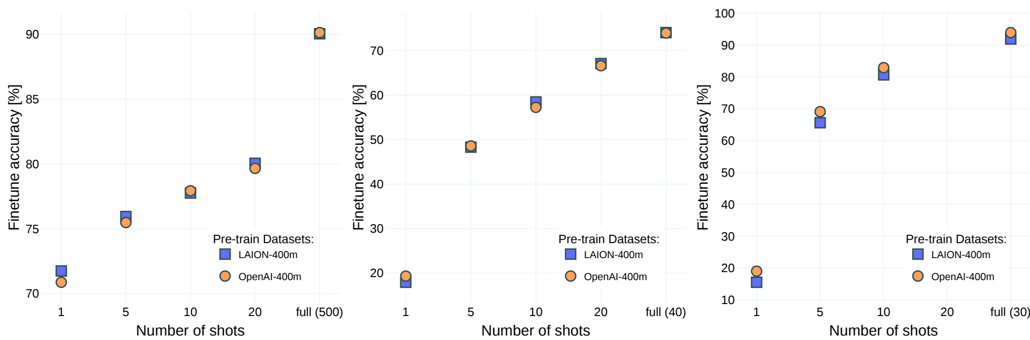
- IN1K-Template-Captions: This dataset includes all data in the ImageNet dataset, paired with templated captions, e.g., "a photo of a classname". This allows us to use CLIP pre-training but on clean images and text. In terms of ImageNet accuracy, this training scheme is very similar to standard supervised training. However, this is now a controlled experiment as we are always using CLIP pre-training.

F Other architectures

In order to see the effect of architecture on the observed trends, we extend the results to the effect of pre-training distribution in Figure 6 to include Vison Transformers. To do so, we used ViT-B/32 released checkpoints trained on LAION-400m and OpenAI-400m, . Figure 8 shows the effect of data distribution on finetune transfer to CIFAR100, DTD, and CALTECH101 when using ViT instead of ResNet-50. While similar to Figure 6 the difference between the fine-tune performance is minimal, we observe that both models perform also very similarly in the few-shot setting. We hypothesize that this observation could be attributed to the similarity between LAION and OpenAI distributions rather than employing a transformer instead of ResNet-50. A controlled study may include to replicate Figure 6 but with ViT, and we leave that for future work.

G Effect of pre-training data distribution: SimCLR instead of CLIP

In contrast to previous experiments with CLIP where we fine-tuned end-to-end from the zero-shot pre-trained model, in SimCLR finetuning we fine-tune using LP-FT (39) because we are no longer able to start with a zero-shot pre-trained model. When we compare to CLIP, we fine-tune both models with LP-FT to facilitate a fair comparison. LP-FT is the following two-step procedure: for each number of shots k we first freeze the encoder and train a classification head from random initialization using k examples per-class from the downstream task. In the second step, we initialize the classification head with this linear probe (LP) then unfreeze all weights and finetune (FT) the whole model.



(a) CIFAR100 (b) DTD (c) CALTECH101

Figure 8: **Effect of the pre-training data distribution: ViT instead of ResNet-50** While similar to Figure 6 the difference between the fine-tune performance is minimal, we observe that both models perform also very similarly in the few-shot setting. We hypothesize that this observation could be attributed to the similarity between LAION and OpenAI distributions rather than employing transformer instead of ResNet-50.

H Extended related works

Transfer learning is widely used in deep learning research and practice and has become a cornerstone in both computer vision and natural language processing. Through the years, there have been many questions on why transfer helps and how to choose a good pre-trained model to transfer from. **(author?)** (40) separated the effect of feature reuse from that of learning low-level pre-training data statistics. **(author?)** (41) investigate the similarity of the pre-training and downstream datasets by looking into medical datasets and found that transfer learning from ImageNet pre-trained models shows little benefit in performance. **(author?)** (42) studied the downstream performance of self-supervised models and found that the best self-supervised models of that time could outperform supervised pre-training as an upstream source of knowledge transfer and that the performance of self-supervised models on ImageNet is indicative of downstream performance on natural image classification tasks. Similarly, **(author?)** (43) found that contrastively trained models consistently outperform standard cross-entropy models in transfer learning. **(author?)** (44) showed that self-supervised models outperform supervised models on ImageNet, even when trained on random and uncurated images from the web. Moreover, they showed that these models are also good at few shot learning by achieving 77.9 % top-1 accuracy using only 10 % on ImageNet.

Building on contrastive techniques, **(author?)** (1) introduced CLIP which learns a joint embedding space for both images and their descriptive captions, making it possible to effectively leverage a large-scale dataset from the Internet. Flamingo (3), a visual language model, is another successful example in the line of multimodal models and enables visual question answering and image captioning. CLIP and similar models like ALIGN (45), BASIC (2), and LiT (46) demonstrated unprecedented robustness to challenging data distribution shifts. This accomplishment raised questions on the probable sources of such robustness—whether this robustness is caused by language supervision, the pre-training data distribution, size, or contrastive loss functions.

(author?) (38) investigated this question and found that the diverse training distribution is the main cause of the robustness properties of CLIP. **(author?)** (47) explored the role of the pre-training dataset for CLIP with a testbed of six pre-training sources, finding that no single pre-training dataset consistently performs best. In recent work, **(author?)** (48) carefully investigated the effect of language supervision in CLIP-like models, finding it an important factor if the pre-training dataset is large and the captions are descriptive enough. Unlike their work, we consider end-to-end fine-tuning which result in higher accuracy.

I Datasets

I.1 Downstream tasks

We have used 9 different downstream datasets. tab:downstreamdatasets describes the first six datasets in Figure 6. While these six datasets are internet-crawled datasets and are more common in transfer learning in computer vision benchmarks, we include three new downstream datasets that are domain-specific, *i.e.*, the dataset is created after a specific challenge is defined in a specific domain.

- EuroSAT (22): The task is to classify land use and land cover based on Sentinel-2 satellite images. The dataset covers 13 spectral bands and consists of 10 classes within a total of 27,019 labeled and geo-referenced images. we create an 80%-20% random class-balanced split with the provided dataset.
- Cassava Leaf Disease Classification (23): The dataset contains 21,397 images from the Kaggle competition, to give farmers access to methods for diagnosing plant diseases. The images are labeled as healthy or as one of four different diseases. we split the dataset with 80%-20% random class-balanced ratio for train and test, respectively.
- Caltech Camera Traps-20 (24): CCT-20 contains 57,864 images in 15 classes, taken from camera traps deployed to monitor animal populations. Classes are either single species *e.g.*, "Coyote") or groups of species, *e.g.*, "Bird"). CCT-20 is a subset of the iWildCam Challenge 2018, whose yearly editions have been hosted on Kaggle. Here we study the subset of CCT-20 that come from the same locations [2], including 14,071 and 16,395 images for train and test respectively.

Table 1: Details on the downstream datasets used in the experiments.

width=1

Downstream Task	Description
CIFAR100	The task consists in classifying natural images (100 classes, with 500 training images each). Some examples include apples, bottles, dinosaurs, and bicycles. The image size is 32x32.
DTD	The task consists in classifying images of textural patterns (47 classes, with 120 training images each). Some of the textures are banded, bubbly, meshed, lined, or porous. The image size ranges between 300x300 and 640x640 pixels.
CALTECH-101	The task consists in classifying images of objects (9144 images in 101 classes plus a background clutter class), including animals, airplanes, chairs, or scissors. The image size varies, but it typically ranges from 200-300 pixels per edge.
PETS	The task consists in classifying images of cat and dog breeds (7000 images in 37 classes). Images dimensions are typically 200 pixels or larger
REAL	The task is a subset of larger DomainNet from six distinct domains, including photos (real), painting, clipart, quickdraw, infograph, and sketch. Total size of 172,000
CLIPART	The task is a subset of larger DomainNet from six distinct domains, including photos (real), painting, clipart, quickdraw, infograph, and sketch. Total size of 172,000

I.2 Pre-training datasets

Our study covers 7 pre-training datasets as follow:

- YFCC: Our experiments mostly include YFCC-2.7M, a random subset of YFCC-15M. The 15M subset of the YFCC-100M dataset (10) was filtered to only include images with English titles or descriptions. The dataset contains 14,829,396 images with natural language captions associated with each image. The images and captions are collected from Flickr.

[2]"Cis" in the main dataset refers to images from locations seen during training, and "trans" refers to new locations not seen during training

- LAION (11): The images and corresponding alt-texts come from web pages collected by Common Crawl (49) between 2014 and 2021. We randomly select a subset of 2.7M and 15M samples for our experiments.
- Redcaps (12): Redcaps contains 11,882,403 examples from 350 manually curated subreddit collected between 2008 and 2020. The subreddits are selected to contain a large number of image posts that are mostly photographs and not images of people.
- Shutterstock: 11,800,000 images and captions from the Shutterstock website.
- Conceptual Captions-3m (13): The raw descriptions in Conceptual Captions are harvested from the alt-text HTML attribute associated with web images. This dataset contains 2,799,553 samples, denoted as CC_2.7m in the plots.
- Conceptual Captions-12m (14): A dataset with 12 million image-text pairs. It is larger than CC_2.7m and covers a much more diverse set of visual concepts. We randomly select 2.7M samples from this dataset, denoted as CC_12_2.7m.
- WIT (15): Image-text pairs come from Wikipedia pages. We use reference description as the source of text data and obtain 5,038,295 examples in total after filtering to include only the English language.

tab:pretraindatasets shows their main source and total size. We also show some examples of image-caption pairs randomly selected from Shutterstock in Figure 9, Redcaps in Figure 10, YFCC-15m in Figure 11, LAION-15m in Figure 12, Conceptual Captions in Figure 13, and WIT in Figure 14. tab:top20$_{w}ordsalsoshowsthemostcommonwordsincaptionsofthesepre-trainingdatasets.$

Looking at Redcaps samples in Figure 10 and also the top 20 captions shows many samples of animals. This is showing why Redcaps perform better on PETS. Samples from WIT in Figure 14 and also its top 20 words mostly featuring geographical locations, which is rare in our downstream task, hence performing worst compared to other pre-training distributions. Shutterstock top 20 words also include words like ”pattern”, ”texture”, ”and design” which are close to DTD classes, hence showing superior performance in this downstream task.

Table 2: Details on pre-training datasets

Dataset	Source	Total size
YFCC	Flickr	14,826,000
LAION	Common Crawl	15,504,742
CC-12M	Unspecified web pages	9,594,338
RedCaps	Reddit	11,882,403
WIT	Wikipedia	5,038,295
ShutterStock	ShutterStock	11,800,000
IN1K-Captions	ImageNet	463,622

Table 3: Most common words in captions of pre-training distributions

width=1

Pre-training dataset	Top 20 words in 1M sample of captions
Shutterstock	**background**, vector, illustration, **design**, icon, **pattern**, **texture**, style, woman, concept, hand, color, flower, view, template, line, business, logo, card, symbol
Redcaps	day, today, year, time, cat, plant, friend, anyone, picture, baby, guy, week, dog, home, morning, night, month, way, boy, work
YFCC-15m	photo, day, park, street, city, picture, view, time, world, year, house, state, center, part, garden, shot, image, building, road, museum
LAION-15m	photo, stock, image, black, woman, design, set, vector, white, print, home, men, blue, dress, art, card, sale, gold, bag, cover
CC-12m	illustration, stock, art, design, photo, image, background, room, vector, house, home, woman, wedding, style, photography, royalty, car, fashion, girl, world
CC-3m	background, actor, artist, player, illustration, view, woman, man, football, team, tree, premiere, city, vector, day, girl, beach, game, hand, people
WIT	view, church, station, map, house, building, hall, museum, city, location, street, park, river, state, john, county, town, center, bridge, world

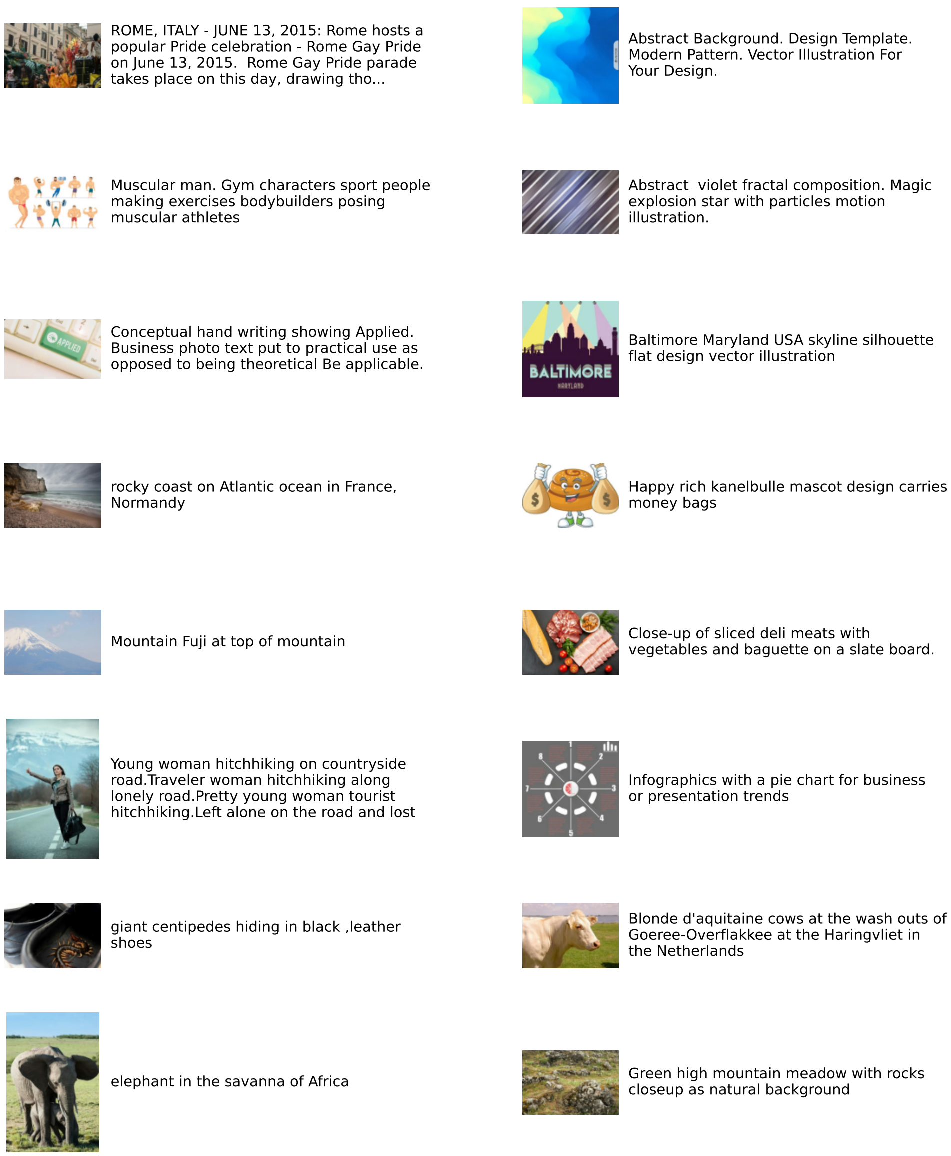


Figure 9: **Random training samples from Shutterstock**

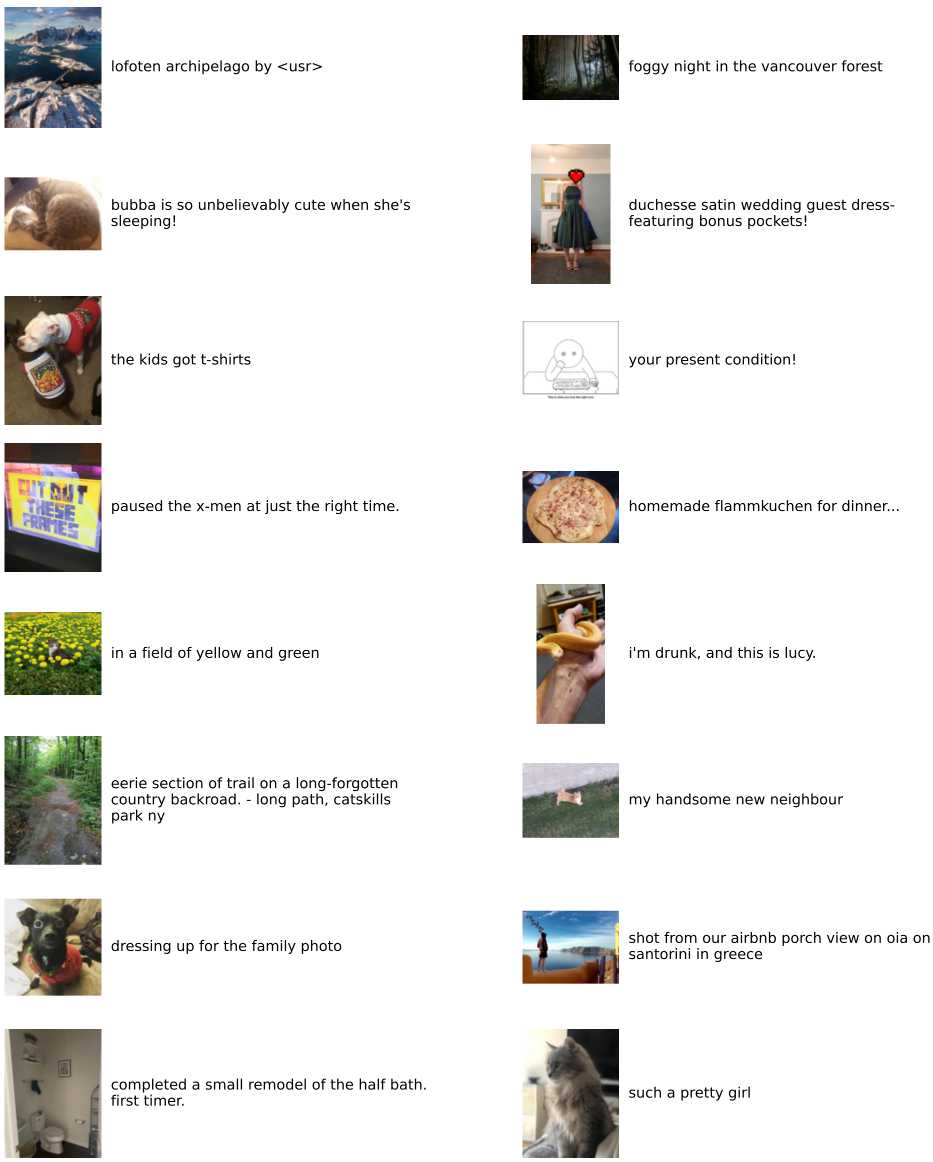


Figure 10: **Random training samples from Redcaps**

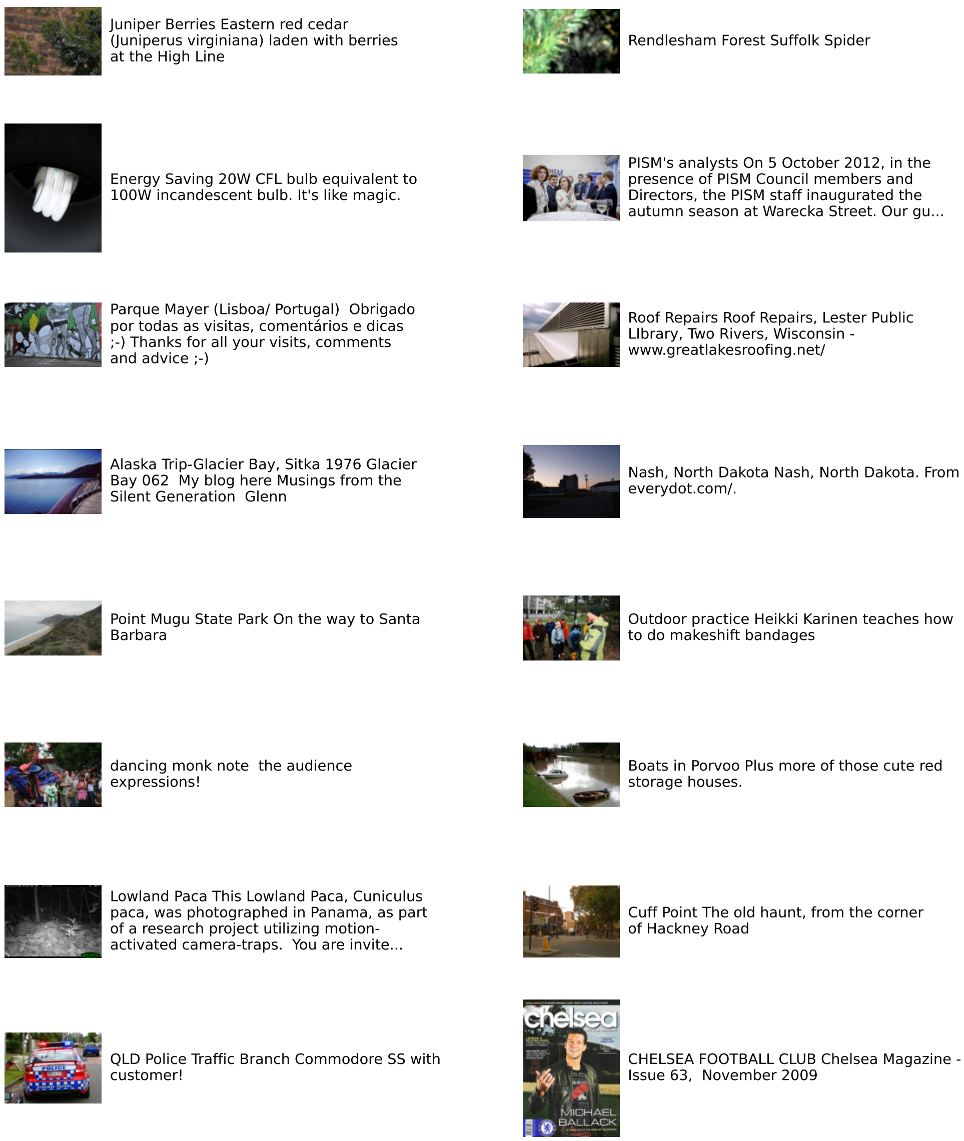


Figure 11: **Random training samples from YFCC**

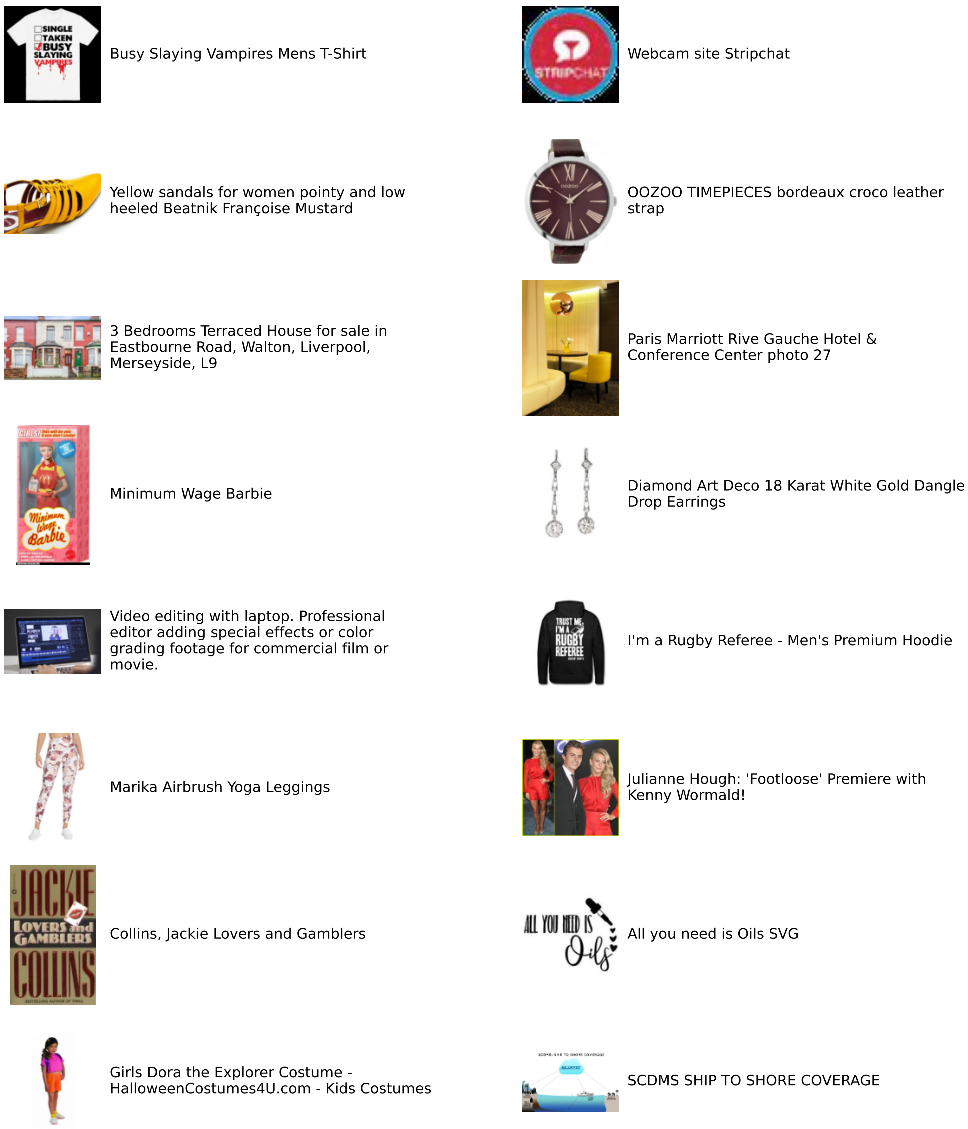
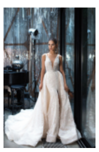


Figure 12: **Random training samples from LAION**

<PERSON> `` The wolf and the lamb shall feed together, and the lion shall eat straw like the bullock: and dust shall be the serpent's meat. They shall not hurt no...

Biker girl in a leather jacket on a black and red color motorcycle

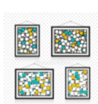


Today's wedding dress inspiration brings us fabulous bridal gowns from creative designer <PERSON>. The Divine Affection lastest bridal collection of <PERSON> wedd...

Easy Cabbage Rolls that are <PERSON>, <PERSON> and have no rice! <PERSON> budget friendly comfort food recipe adapted from my Russian grandmother!

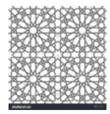
Wedding rings on a bouquet of roses stock photos

All Balls Swinging Arm Bearing Kit for Yamaha XT225 | XT250 Serow 1993 to 2007



Different types of photo frames with circles and squares on the wall - background template stock illustration

The art of good drinking



Islamic vector geometric ornaments based on traditional arabic art. Oriental seamless pattern. Muslim mosaic. Turkish, Arabian tile on a white background. Mosque ...

Light Touch Wall digital marketing activation at the Canberra Centre.

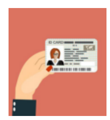


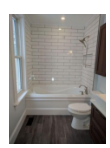
Illustration of hand holding the id card. Vector illustration flat design.

<PERSON>: U. <PERSON> in United States Army. First <PERSON> appointed to that position. First, &, so far, only <PERSON> to serve on Joint Chiefs of Staff. Black H...

<PERSON> tattoo, the American number 23 from Akron, United States

Search the hidden word, the simple educational kid game. stock illustration

The Russian army entering Prussia, 1914 : News Photo



Modern Bathroom Makeovers 20 Design Ideas For a Small Bathroom Remodel. Modern Bathroom Designs On A Budget Minimalist Small Bathrooms, Modern Small Bathrooms, Mo...

Figure 13: **Random training samples from Conceptual Captions**


Japanese chamber pot from the Edo period


Ratel

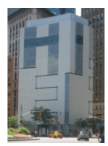

15th race in 1982


Dedication plaque at Oregon Dunes Overlook, Oregon Dunes National Recreation Area

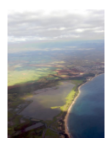

Robinson in March 2018


85-15 Wareham Place, Donald Trump's childhood home


Hugo van der Goes, Saint Luke Drawing the Virgin, c. 1470–80. National Museum of Ancient Art, Lisbon


Museum of Arts & Design at 2 Columbus Circle, nearly completed in July 2008. A piece by David Dunlap's in the NY Times reveals that the appearance of the letter ...


Paul McAuley at Worldcon 2005 in Glasgow


Maui Veterans Highway shown just above Kealia Pond National Wildlife Refuge as it enters Kihei


Construction of restrooms and locker rooms with east side stands and pavilion


McLaren 600LT


Buzz Aldrin received praise for his performance.


James Barber House


A Fiat G.91PAN, in service in the Frecce Tricolori from 1963 to 1982


Alexa Stirling, c. 1919

Figure 14: **Random training samples from WIT**