# Selectivity Drives Productivity: Efficient Dataset Pruning for Enhanced Transfer Learning

Yihua Zhang[1*], Yimeng Zhang[1*], Aochuan Chen[1*], Jinghan Jia[1], Jiancheng Liu[1], Gaowen Liu[2], Mingyi Hong[3], Shiyu Chang[4], Sijia Liu[1]
[1]Michigan State University, [2]Cisco Research, [3]University of Minnesota, Twin Cities, [4]University of California, Santa Barbara

Paper | Code | Website

## Dataset Pruning for Transfer Learning

❖ Not all sources classes are necessary or beneficial[1]
- Some source data could be **harmful** to downstream performance.
- **Removing** specific source classes can **improve** transfer learning.

❖ Conventional dataset pruning lacks effectiveness on transfer learning.
- Conventional **SOTA DP** methods do **NOT** yield significant improvement over **random** pruning on transfer learning.

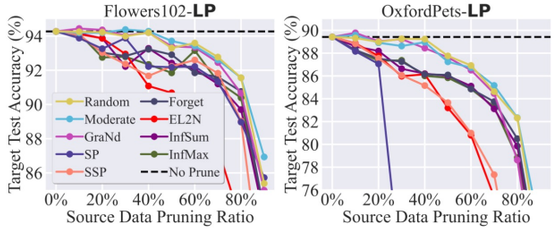
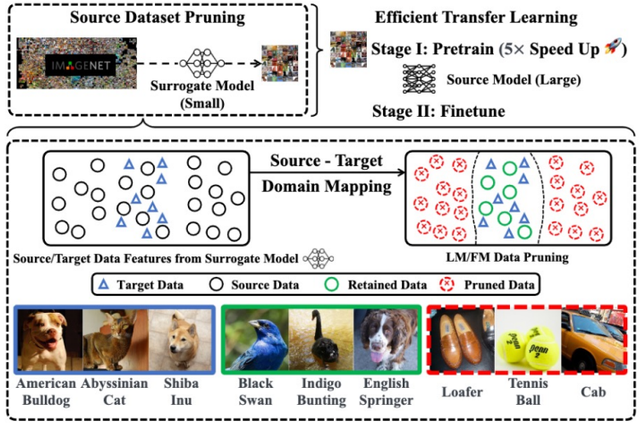



**Figure 1.** Transfer learning accuracy of existing DP methods on ImageNet at different pruning ratios, where ResNet-101 is the source model.

## An Overview of Our Proposal



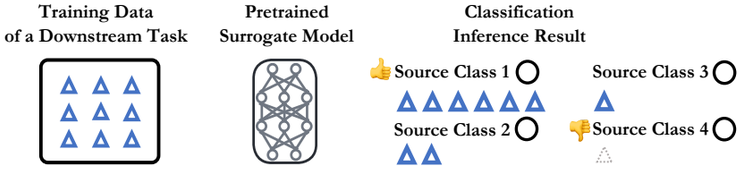

## Research Question

- Existing method: brute-force based, effective but not affordable.

**(Q) How can we extend DP to transfer learning with high efficiency, broad applicability, and lossless or even improved target performance?**

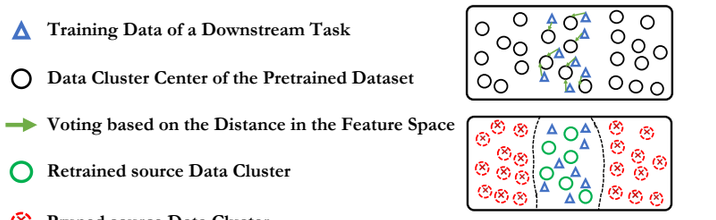
## Label Mapping for Supervised Pretraining

- **Rationale behind the design**: source data **similar** to downstream data intend to contribute more during the transfer process.
- **Dataset selection as a voting process**: each downstream training data can vote for its most similar/relevant source training class.
- **Label mapping** through a pretrained small surrogate model.




## Feature Mapping for Self-Supervised Pretraining

- **Data cluster as the basic pruning unit when labels are unavailable.**




*[1] S. Jain et al. "A data-based perspective on transfer learning." CVPR 2023.*
*[2] S. Kim et al. "Coreset sampling from open-set for fine-grained self-supervised learning " CVPR 2023*

## Experiment Result Highlights

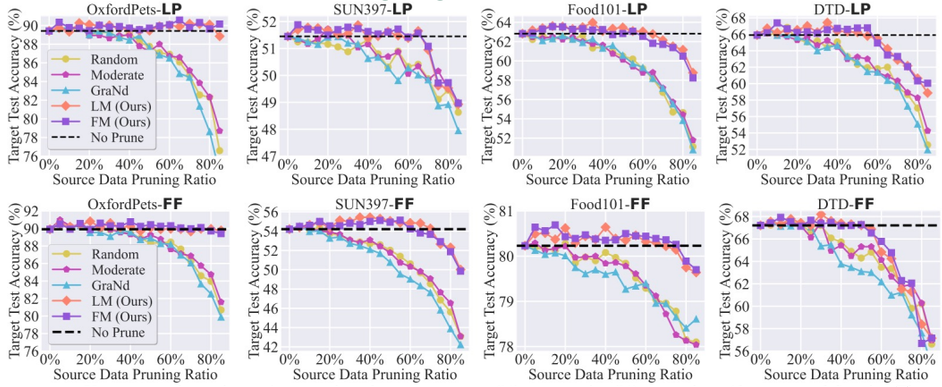



**Figure 2.** Unstructured pruning trajectory given by test accuracy (%) vs. sparsity (%) on various (dataset, model) pairs. The performance of dense model and that of the best winning ticket are marked using dashed lines in each plot. The solid line and shaded area of each pruning method represent the mean and variance of test accuracies over 3 trials.

| Dataset | OxfordPets | | | | | SUN397 | | | | | Flowers102 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pruning Ratio | 0% | 50% | 60% | 70% | 80% | 0% | 50% | 60% | 70% | 80% | 0% | 50% | 60% | 70% | 80% |
| RANDOM | 69.26 | 62.32 | 61.27 | 59.09 | 53.75 | 47.36 | 45.63 | 45.08 | 43.54 | 39.81 | 85.17 | 82.23 | 82.60 | 81.03 | 80.02 |
| MODERATE | | 63.37 | 62.45 | 63.31 | 57.42 | | 45.73 | 45.14 | 44.23 | 40.82 | | 82.45 | 81.45 | 81.69 | 81.32 |
| GRAND | | 64.42 | 63.34 | 61.14 | 56.42 | | 45.72 | 45.58 | 45.24 | 41.72 | | 82.85 | 82.44 | 82.14 | 81.73 |
| FM (ours) | | **69.92** | **69.99** | **70.29** | **70.21** | | **48.46** | **48.58** | **47.90** | **46.00** | | **85.22** | **85.42** | **84.37** | **84.61** |

**Table 1.** The downstream performance with different source data pruning ratios in the SSL pretraining setting. A randomly initialized RN-101 is self-supervised pretrained using MOCO V2 on each full/pruned source dataset and finetuned on the downstream task through LP.

| Pruning Ratio | 0% | 20% | 40% | 60% | 80% |
|---|---|---|---|---|---|
| Time Consumption (h) | 5.4 | 4.6 (15%↓) | 3.5 (35%↓) | 2.4 (56%↓) | 1.3 (76%↓) |

**Table 2.** Time consumption of LM/FM to obtain the pretrained model. The reported time consumption covers surrogate model (RN-18) training, LM/FM dataset pruning, and source model pretraining (RN-101).

| Method | SUN397 | | | | DTD | | | |
|---|---|---|---|---|---|---|---|---|
| | Pruning Ratio | Acc.(%) LP | Acc.(%) FF | Time (h) | Pruning Ratio | Acc.(%) LP | Acc.(%) FF | Time (h) |
| DENSE | N/A | 51.45 | 54.21 | 5.4 | N/A | 65.91 | 67.21 | 5.4 |
| DENSE-ADV | N/A | 52.97 | 55.67 | 13.7 | N/A | 67.23 | 68.92 | 13.7 |
| LM | 70 | 50.95 | 54.28 | 1.9 | 50 | 66.25 | 67.22 | 2.9 |
| LM-ADV | 70 | 52.07 | 55.49 | 4.2 | 50 | 67.02 | 68.54 | 6.7 |

**Table 4.** Downstream performance of models pretrained on full/pruned source dataset (Dense/LM) w/wo adversarial pretraining (Adv).

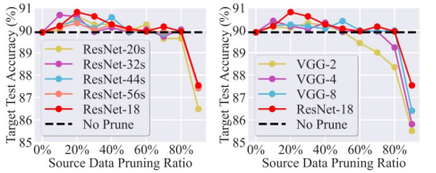



**Figure 3.** Source dataset pruning trajectory given the downstream task OxfordPets using different surrogate models.

**Acknowledgement.** The work of Y. Zhang, Y. Zhang, A. Chen, J. Jia, J. Liu, and S. Liu was supported in part by the Cisco Research Award, the NSF Grant IIS-2207052, and the ARO Award W911NF2310343. The work of S. Chang was supported by the Cisco Research Award and the NSF Grant IIS-2207052. The work of M. Hong was supported by NSF grants CIF-1910385 and EPCN-2311007.

*Contact: zhan1908@msu.edu*