# 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]UC Santa Barbara

[*]Equal contribution

**Paper**

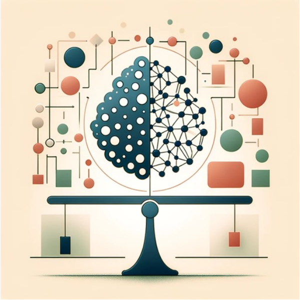
**OPTML Group**

**MSU**

MICHIGAN STATE UNIVERSITY

OPTML

# The Modern Training Paradigm for Big Data



Source Dataset

Target Dataset

Big Data Collection → Model Pretraining → Model Finetuning

## Pretraining-Finetuning Pipeline

MICHIGAN STATE UNIVERSITY

OPTML

# Do We Need All the Source Data?

Recent evidence has shown:

- Some source data could make a **harmful** influence in the downstream performance.

- **Removing** specific source classes can **improve** transfer learning.

# Existing Methods

- Dataset pruning (DP) is a well-studied problem for **in-domain** scenarios:
  - clustering-based methods
  - influence function-based methods
  - training dynamics-based methods
  - …
- DP for **transfer learning** is under-explored
  - Brute force-based method: time-consuming and unaffordable

MICHIGAN STATE UNIVERSITY

OPTML

## Open Question

(DP for transfer learning)
How to efficiently prune source data to obtain a subset, ,
with lossless or improved transfer learning accuracy of
the source model on a target task?

MICHIGAN STATE UNIVERSITY

OPTML

# Conventional DP is NOT Effective for TL!

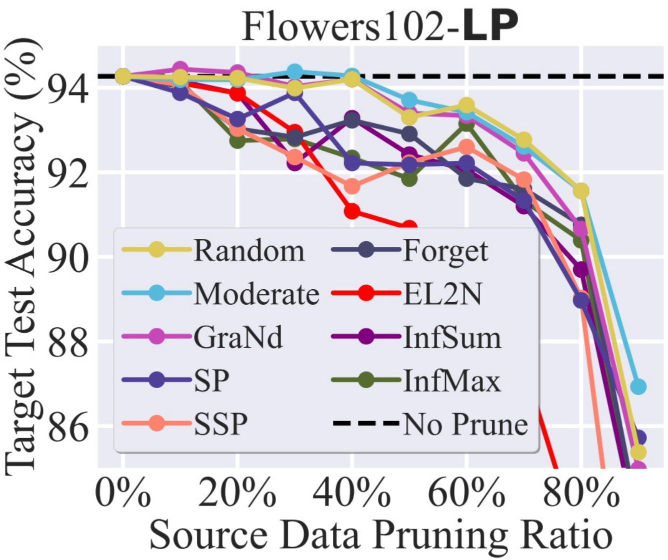



Figure 2: Transfer learning accuracy of existing DP methods on ImageNet at different pruning ratios, where ResNet-101 is the source model

MICHIGAN STATE UNIVERSITY

OPTML

# Conventional DP is NOT Effective for TL!

In transfer learning, conventional SOTA DP methods do **NOT** yield significant performance improvement over random pruning!

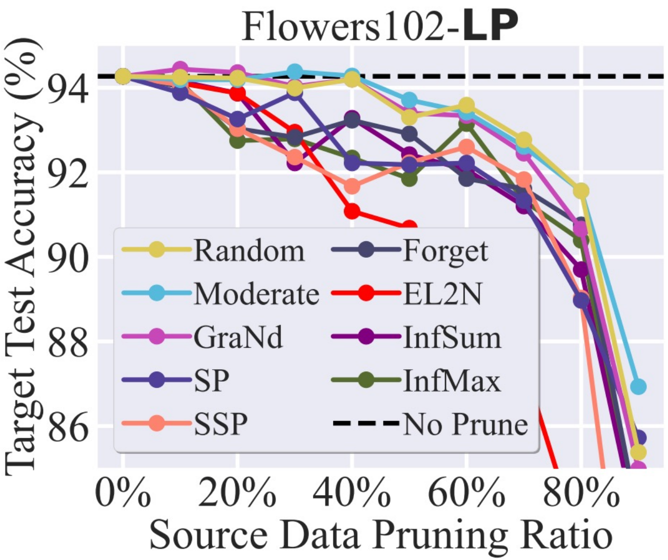



Figure 2: Transfer learning accuracy of existing DP methods on ImageNet at different pruning ratios, where ResNet-101 is the source model

MICHIGAN STATE UNIVERSITY

OPTML

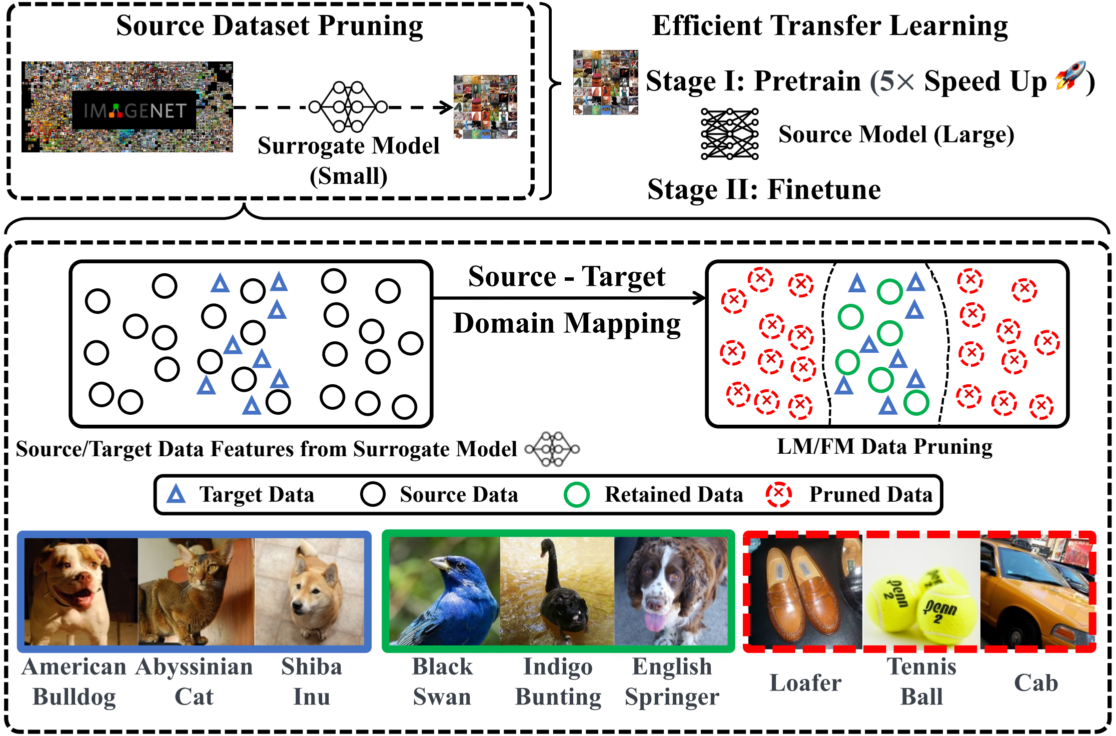
# Our Proposal

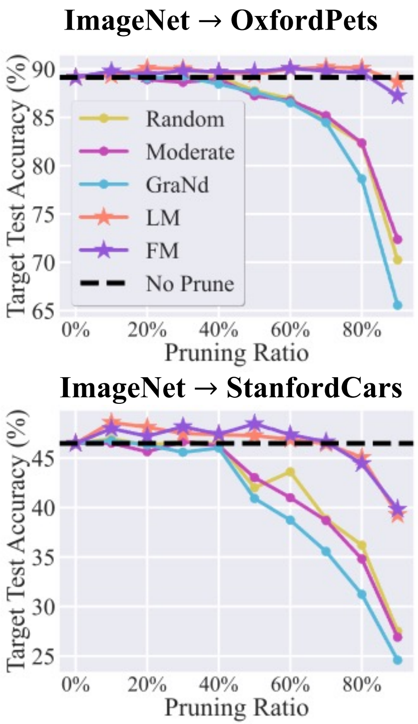
- Rationales behind our proposal:
  - Source data similar to downstream data intend to contribute more during the transfer process;
  - The DP for TL method can be viewed as a “voting” process, each target training data can vote for its most similar source training class;
  - A pretrained small model should help us identify which source class is the most similar to a downstream training data.

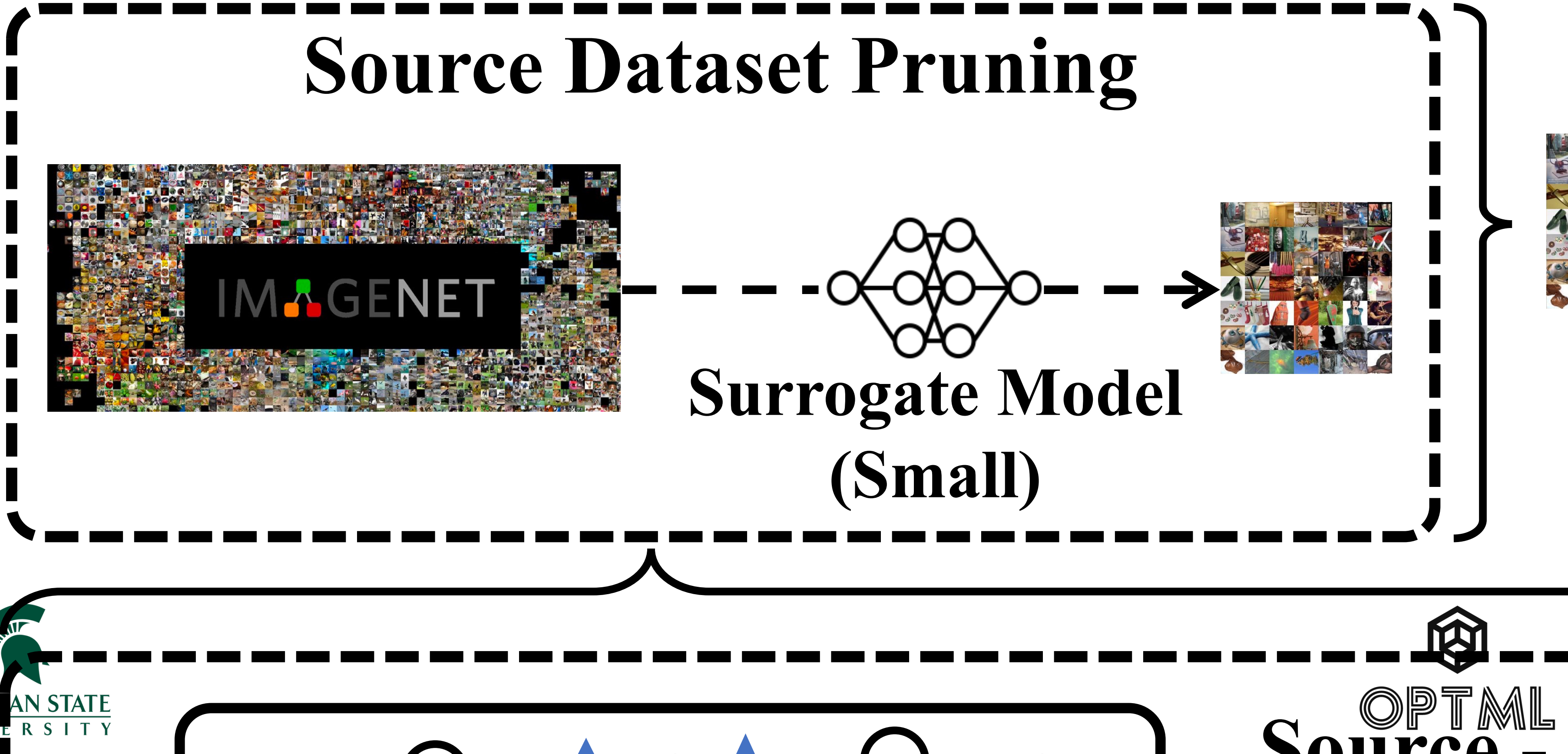
MICHIGAN STATE UNIVERSITY

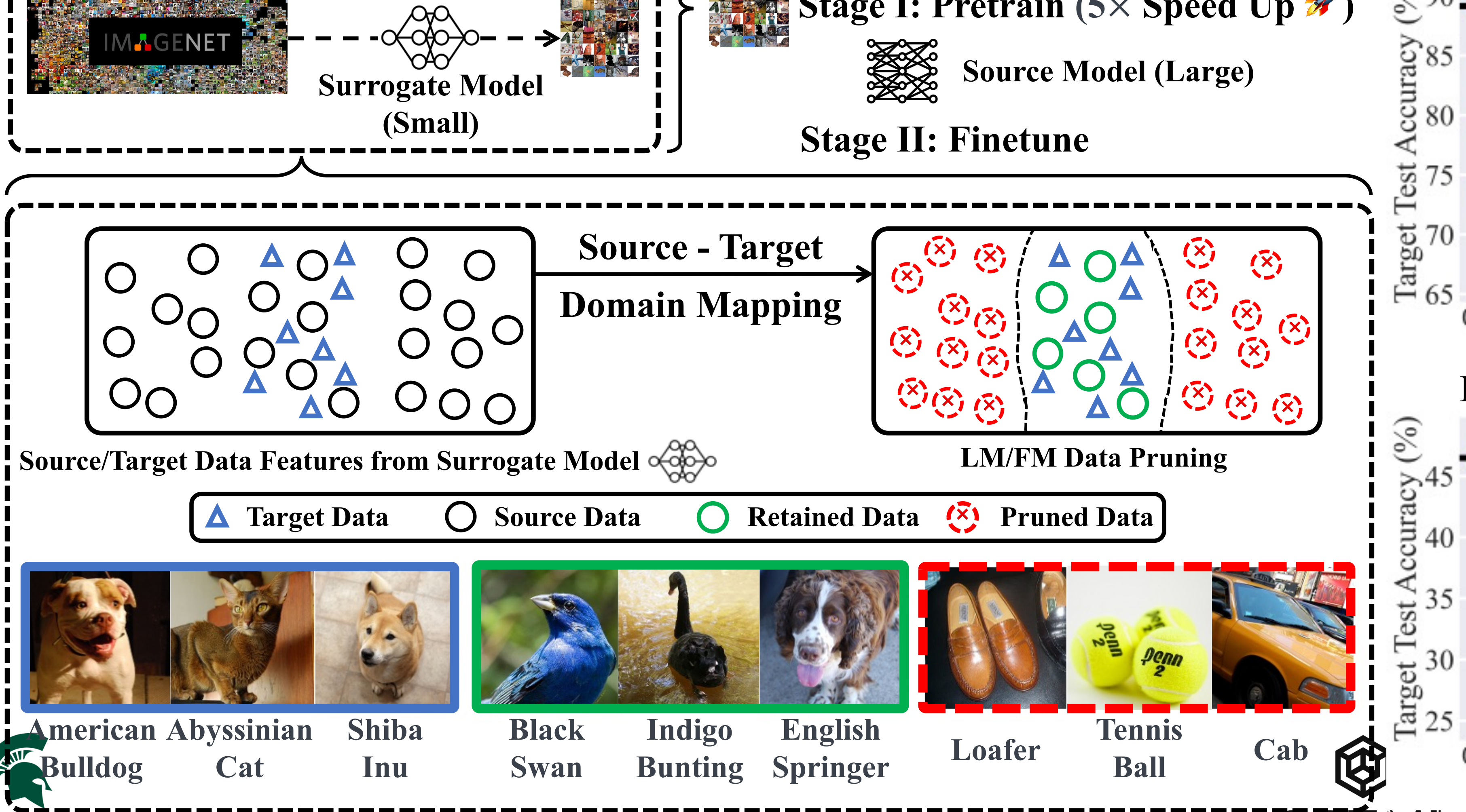
OPTML

# Our Proposal: An Overview



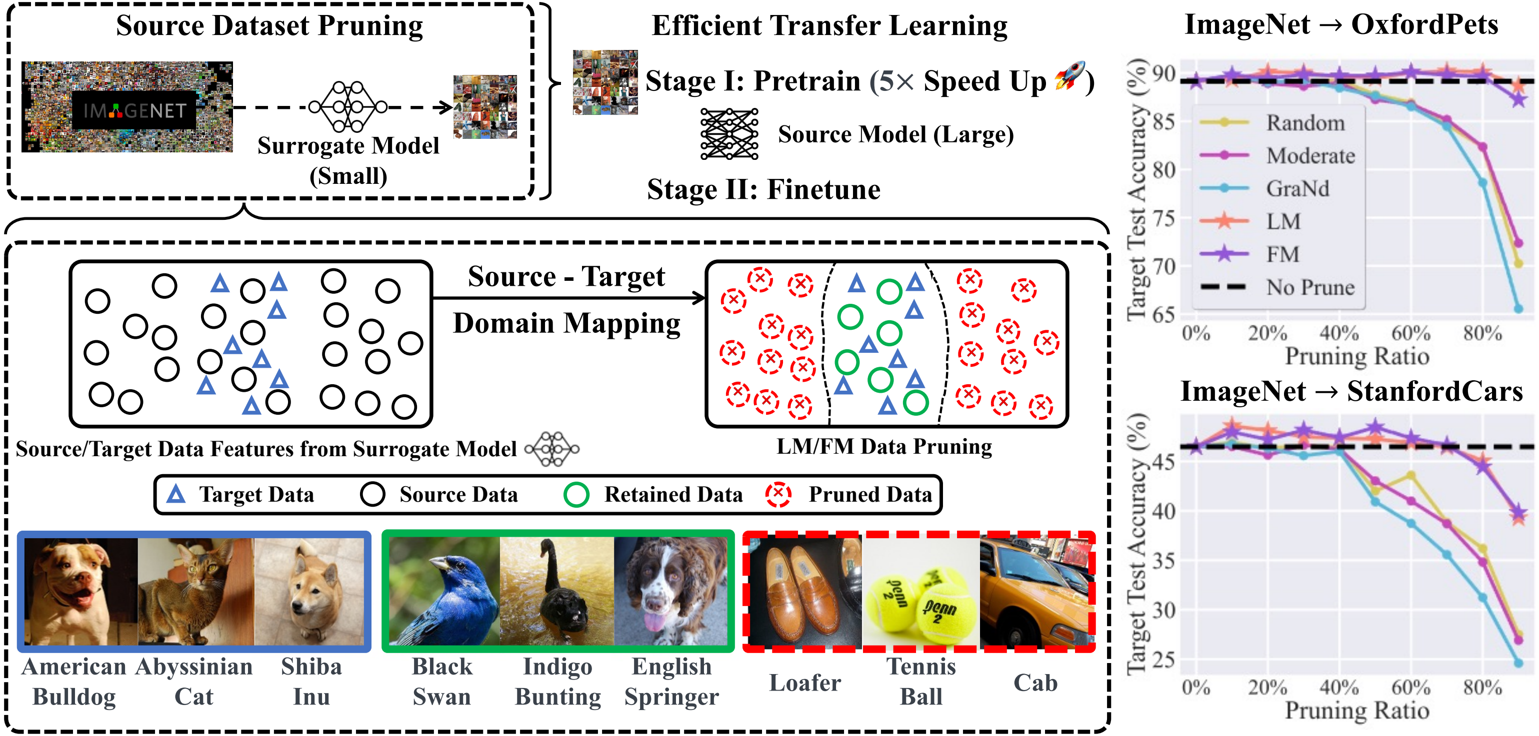




Source Dataset Pruning
IMAGENET
Surrogate Model
(Small)

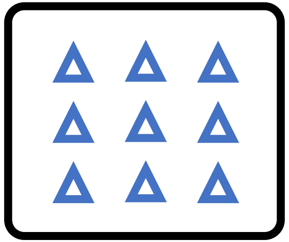


IMAGENET
Surrogate Model
(Small)
Source Model (Large)
Stage II: Finetune
Source - Target
Domain Mapping
Source/Target Data Features from Surrogate Model
LM/FM Data Pruning
Target Data
Source Data
Retained Data
Pruned Data
American Bulldog
Abyssinian Cat
Shiba Inu
Black Swan
Indigo Bunting
English Springer
Loafer
Tennis Ball
Cab
Target Test Accuracy (%)
Target Test Accuracy (%)

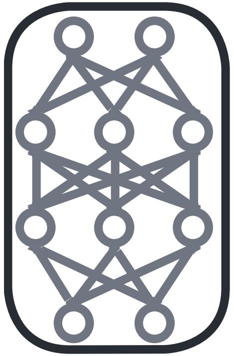
# An Overview


Source Dataset Pruning
IMAGENET
Surrogate Model
(Small)
Efficient Transfer Learning
Stage I: Pretrain (5× Speed Up)
Source Model (Large)
Stage II: Finetune
Source - Target
Domain Mapping
Source/Target Data Features from Surrogate Model
LM/FM Data Pruning
Target Data
Source Data
Retained Data
Pruned Data
American Bulldog
Abyssinian Cat
Shiba Inu
Black Swan
Indigo Bunting
English Springer
Loafer
Tennis Ball
Cab
ImageNet → OxfordPets
Target Test Accuracy (%)
Pruning Ratio
Random
Moderate
GraNd
LM
FM
No Prune
ImageNet → StanfordCars
Target Test Accuracy (%)
Pruning Ratio

Training Data
of a Downstream Task



Pretrained
Surrogate
Model



Classification
Inference
Result

Source Class 1 ○

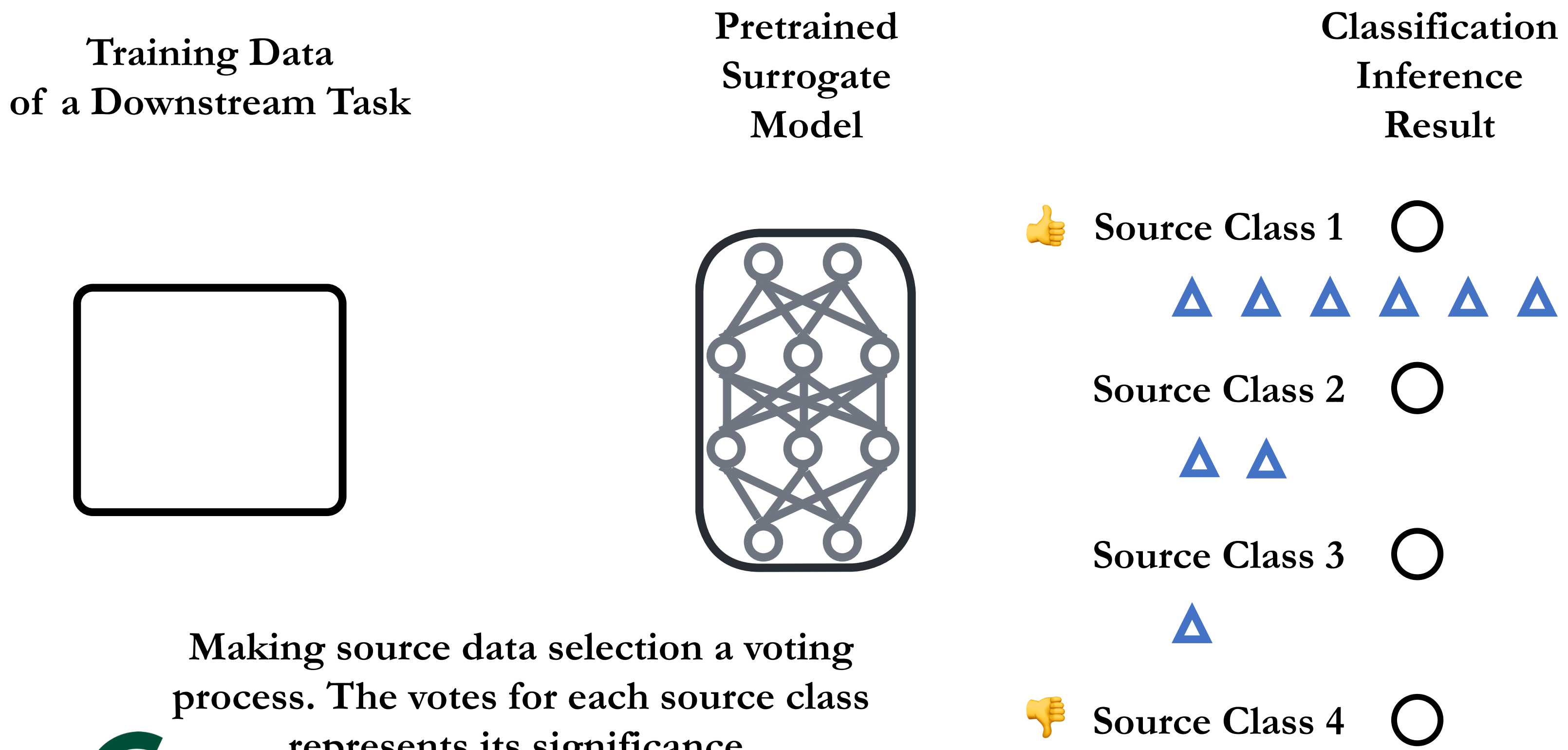
Source Class 2 ○

Source Class 3 ○

Source Class 4 ○

MICHIGAN STATE
UNIVERSITY

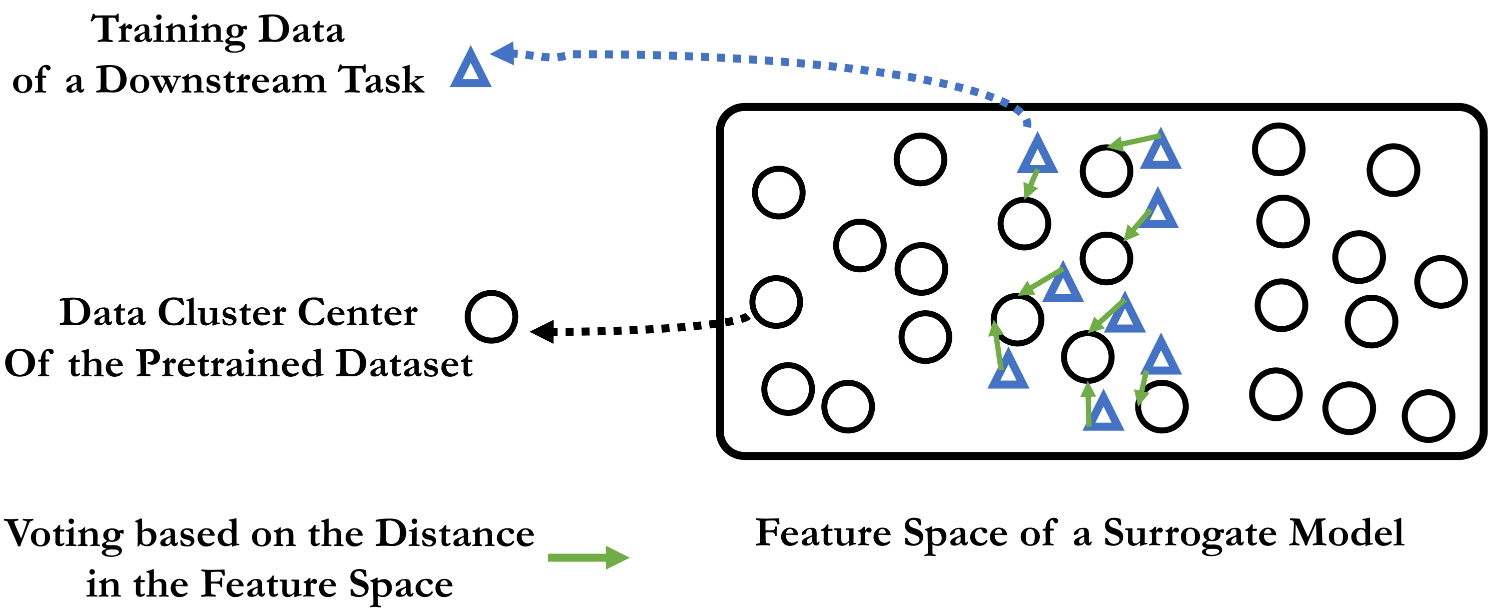
OPTML




Making source data selection a voting process. The votes for each source class represents its significance.

MICHIGAN STATE UNIVERSITY

OPTML

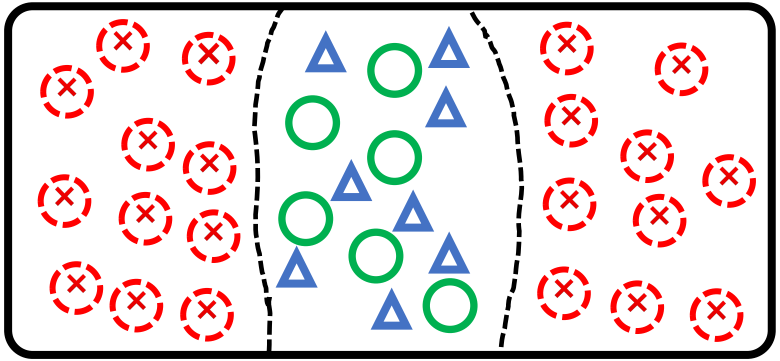

Training Data
of a Downstream Task
Data Cluster Center
Of the Pretrained Dataset
Voting based on the Distance
in the Feature Space
Feature Space of a Surrogate Model

Training Data
of a Downstream Task

Retrained source
Data Cluster

Pruned source
Data Cluster

Voting based on the Distance
in the Feature Space



Feature Space of a Surrogate Model

MICHIGAN STATE
UNIVERSITY

OPTML

# Summary on LM & FM

- Surrogate Model can be very small, or even not well-trained;
- The pruned source dataset can be used for efficiently training much larger models (100× size);
- The model pretrained on the pruned source dataset can be finetuned on the downstream task with lossless or even higher performance;

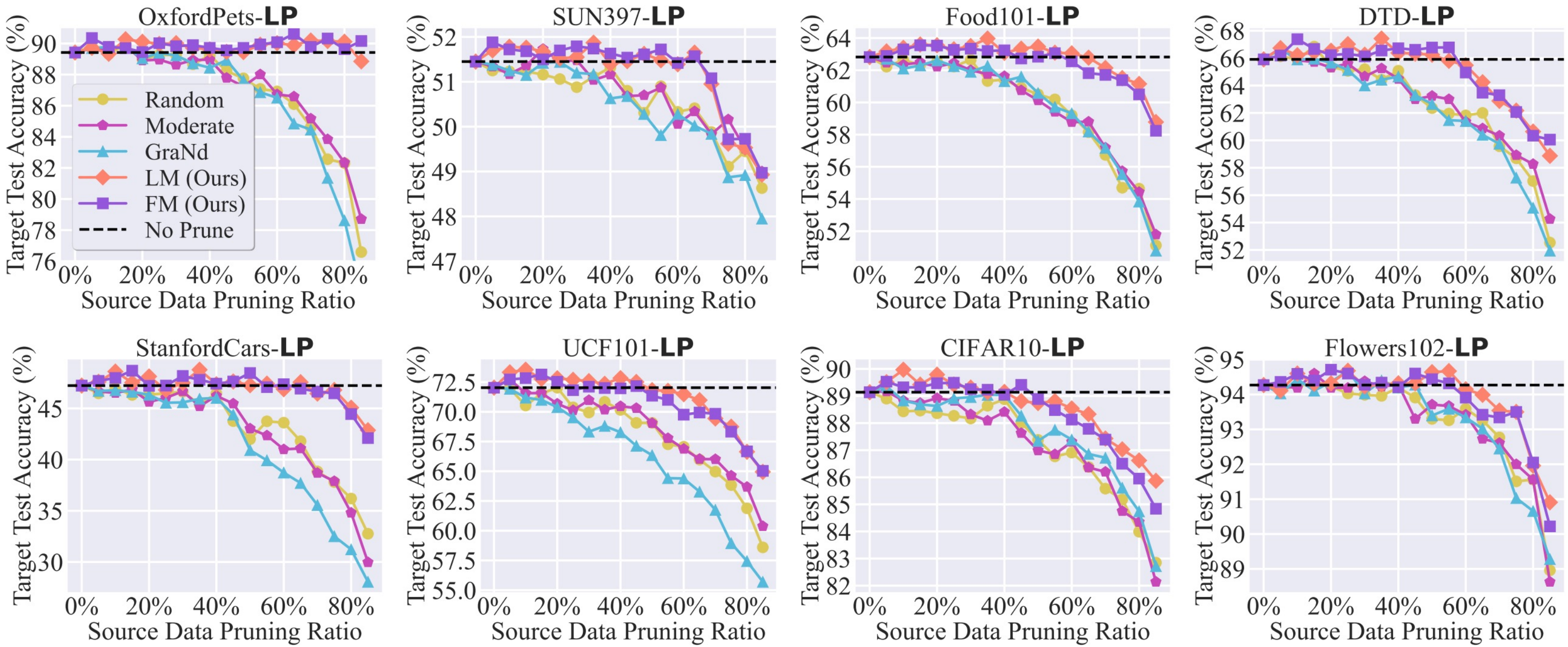
MICHIGAN STATE UNIVERSITY

OPTML

# Experiments Overview

- Transfer learning on 9 datasets on both CNN/ViTs.
- Supervised and Unsupervised methods (MoCo v2/v3);
- DP for adversarial transfer learning;
- Few-shot transfer learning benchmark (VTAB);
- Multi-Task setting;
- Biased-data setting;
- Ablation study:
  - Surrogate model size
  - Reverse order selection
  - Feature distribution analysis

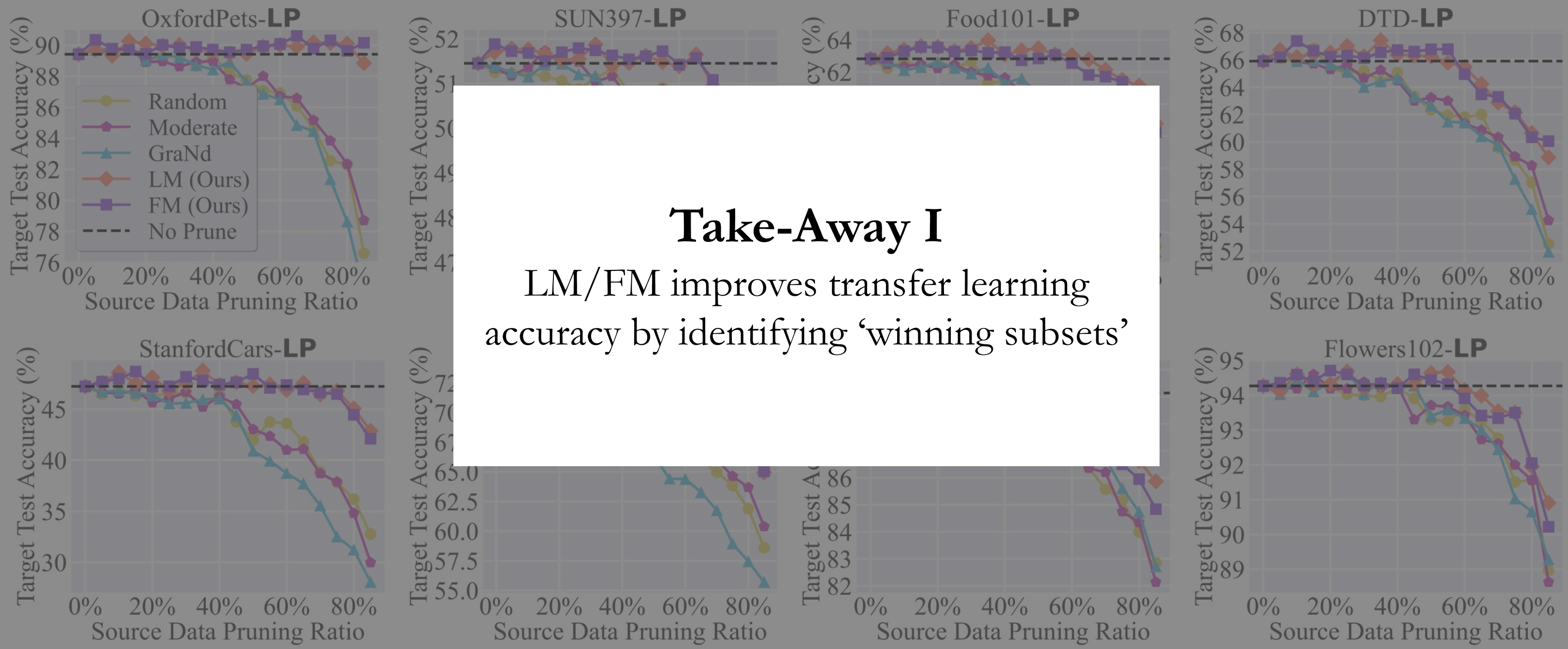
# Results Highlights I




MICHIGAN STATE UNIVERSITY

OPTML

# Results Highlights I




**Take-Away I**

LM/FM improves transfer learning accuracy by identifying 'winning subsets'

MICHIGAN STATE UNIVERSITY

OPTML

# Results Highlights II

Table 2: 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. The best result in each pruning ratio is marked in **bold** and the performance surpassing the unpruned setting (pruning ratio 0%) is highlighted in cyan .

| 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** |

MICHIGAN STATE UNIVERSITY

OPTML

# Results Highlights I

Table 2: The downstream perform the SSL pretraining setting. A
randomly initialized RN-101 is s ch full/pruned source dataset
and finetuned on the downstream g ratio is marked in **bold** and
the performance surpassing the u d in cyan .

| Dataset / Pruning Ratio | OxfordP 0% | 50% | 60% | Flowers102 50% | 60% | 70% | 80% |
|---|---|---|---|---|---|---|---|
| RANDOM | 69.26 | 62.32 | 61.27 | 82.23 | 82.60 | 81.03 | 80.02 |
| MODERATE | | 63.37 | 62.45 | 82.45 | 81.45 | 81.69 | 81.32 |
| GRAND | | 64.42 | 63.34 | 82.85 | 82.44 | 82.14 | 81.73 |
| FM (ours) | | **69.92** | **69.99** | **85.22** | **85.42** | **84.37** | **84.61** |

**Take-Away II**

FM is effective in both supervised and unsupervised transfer learning.

MICHIGAN STATE UNIVERSITY

OPTML

# Results Highlights III

Table 3: Time consumption of LM/FM in Fig.4 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).

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

MICHIGAN STATE UNIVERSITY

OPTML

# Results Highlights I

Table 2: The downstream perform the SSL pretraining setting. A randomly initialized RN-101 is s ch full/pruned source dataset and finetuned on the downstream g ratio is marked in **bold** and the performance surpassing the u d in cyan .

| Dataset / Pruning Ratio | OxfordP 0% | 50% | 60% | Flowers102 50% | 60% | 70% | 80% |
|---|---|---|---|---|---|---|---|
| RANDOM | 69.26 | 62.32 | 61.27 | 82.23 | 82.60 | 81.03 | 80.02 |
| MODERATE | | 63.37 | 62.45 | 82.45 | 81.45 | 81.69 | 81.32 |
| GRAND | | 64.42 | 63.34 | 82.85 | 82.44 | 82.14 | 81.73 |
| FM (ours) | | **69.92** | **69.99** | **85.22** | **85.42** | **84.37** | **84.61** |

**Take-Away III**

FM/LM significantly enhances the training efficiency.

# Potential Applications Related to Lifelong Learning

- Safety/Alignment preservation in transfer learning for large CV models and LLM (large language models).
  - How to perform data selection to preserve safety and alignment gained during pretraining?
  - How to pinpoint the data contributing the most to the general safety/alignment during pretraining?

[Adversarial Robustness: From Self-Supervised Pre-Training to Fine-Tuning]

[Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!]

Paper

terima kasih

mulţumesc

ありがとう

谢谢 ngiyabonga suksema

Met dank

obrigada

Thank

baie dankie

molte grazie

You

merci

감사합니다

obrigado

Danke schön!

謝謝

gracias

Благодарность شكرًا

Спасибі Dziękuję

dank u

mahalo

tusind tak

MICHIGAN STATE UNIVERSITY

OPTML