WED-PM-306:
Large-scale Training Data Search for Object Re-identification

Yue Yao, Tom Gedeon, Liang Zheng

Australian National University, Curtin University

1

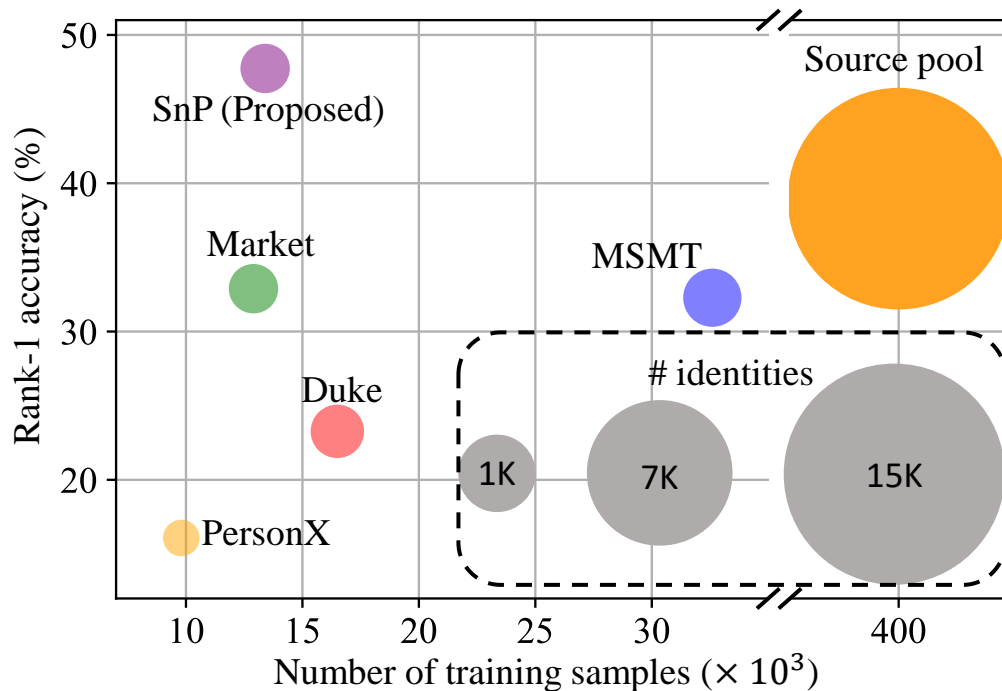# Highlights: training data search from a source pool

Source pool

S — Training set

T — Target validation/test set (unlabeled)

Objective: create a small training set from source pool but can train a model with high accuracy on target data

Highlights: results

When the target is AlicePerson, the searched training set (SnP) has

- higher accuracy than source pool

- higher accuracy than each individual dataset

- much more small scale in the number of samples and IDs

What most works are studying
model-centric

Based on **benchmarks** like ImageNet, COCO, etc

Model Modeling → Training

Constructing Training set

Evaluation on Validation set
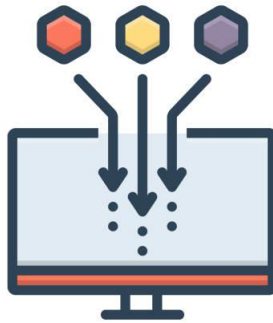
Test / Deployment
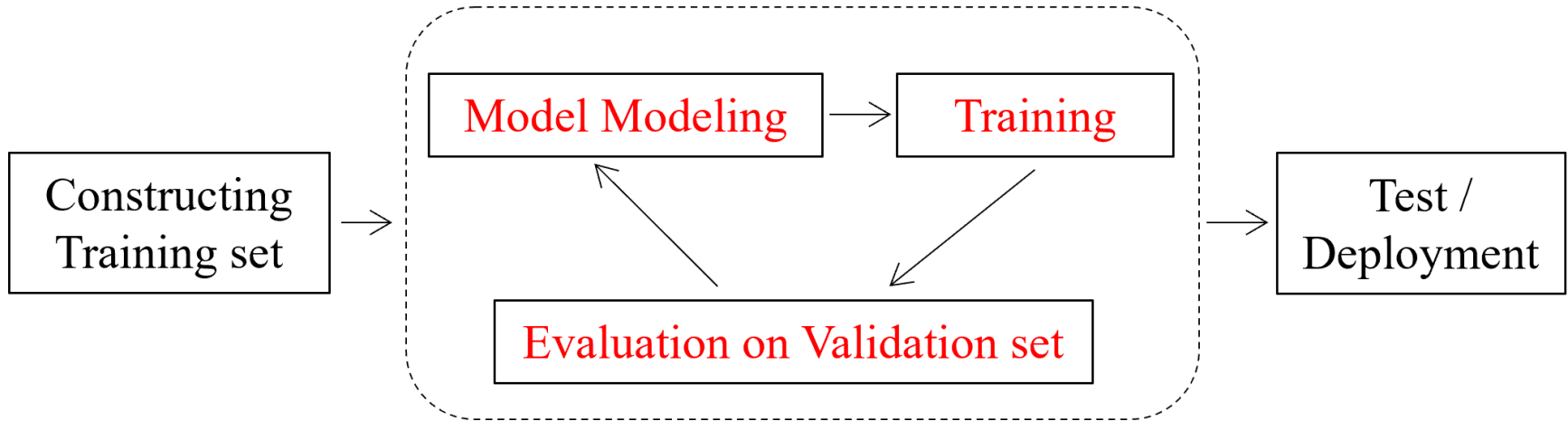
Under a fixed source training and target validation,
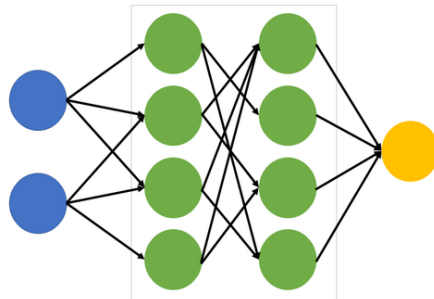Can we improve the training algorithm or model?

4

Suppose we hope to have a deep learning system for a new target
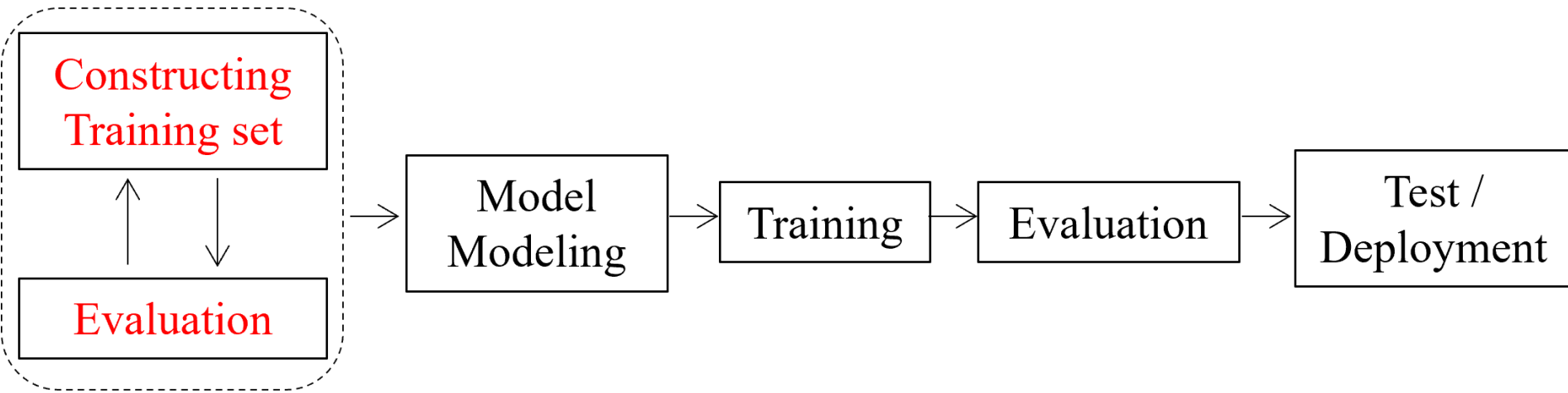
- first half of time collecting/cleaning data

- The other half of time configuring your deep learning network

What I'm going to talk about
data-centric

```
┌ ─ ─ ─ ─ ─ ─ ─ ─ ─ ┐
|  ┌──────────────┐ |
|  | Constructing |                ┌──────────┐      ┌──────────┐    ┌──────────┐              ┌──────────────┐
|  | Training set | |      ──→     |  Model   | ──→  | Training | ──→| Evaluation| ──→         |   Test /     |
|  └──────────────┘ |              | Modeling |      └──────────┘    └──────────┘              | Deployment   |
|     ↑      ↓       |              └──────────┘                                                └──────────────┘
|  ┌──────────────┐ |
|  |  Evaluation  | |
|  └──────────────┘ |
└ ─ ─ ─ ─ ─ ─ ─ ─ ─ ┘
```

Under a fixed target validation,
Can we improve the source training
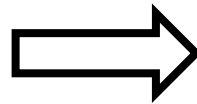data to improve target performance?

Training data search: from a source pool

Source pool

Train directly?

$\implies$

Target validation/test set (unlabeled)

Not a good idea:
- Time costly
- There could be a better choice

Training data search: from a source pool

Source pool

Training set

Target validation/test set (unlabeled)

Objective: create a small training set from source pool but can train a model with high accuracy on target data

We may select training from real training

Source pool

→ Training set A, B, C, D, E

Real training and validation

↓

T

validation set

Data comparison
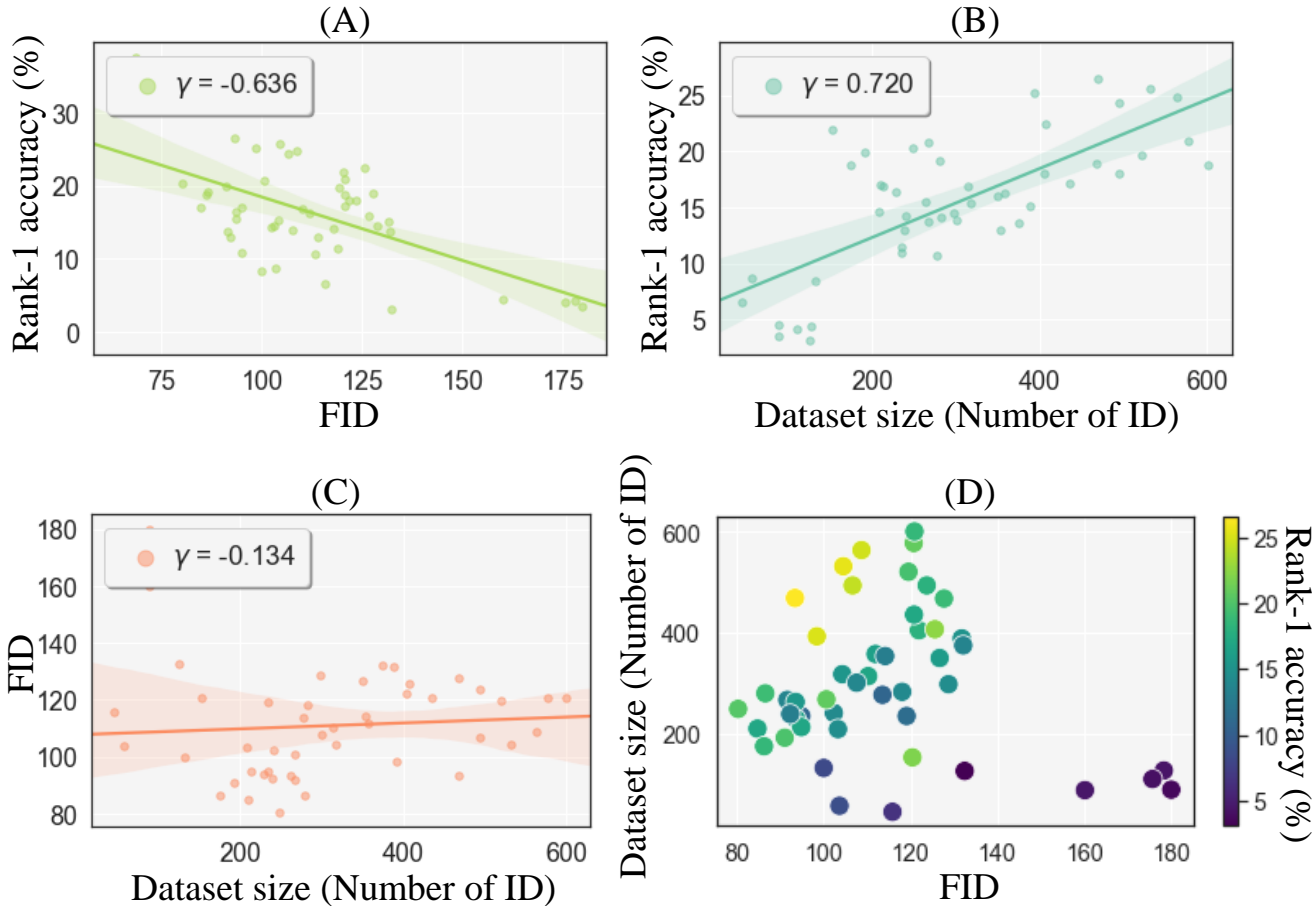
B ≻ D ≻ C ≻ A ≻ E ←

A **time-consuming** process!
Require validation labels!

9

If we can find a good training set without real training

Person re-ID

Negative correlation to the domain gap (FID)
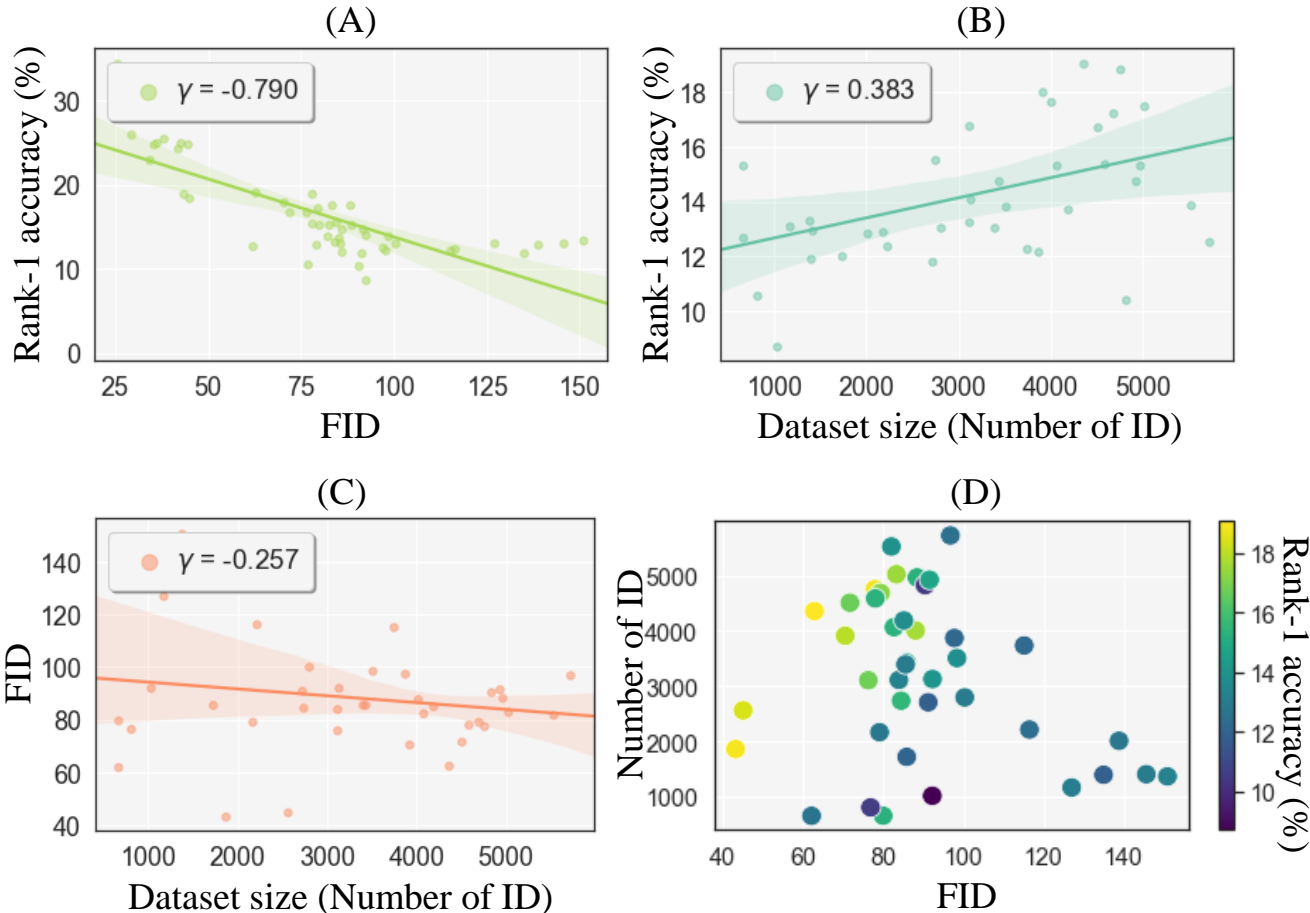
Positive correlation to the dataset size

Jointly achieve best accuracy

Every point is a dataset

More experiments

Vehicle re-ID

(A)

(B)

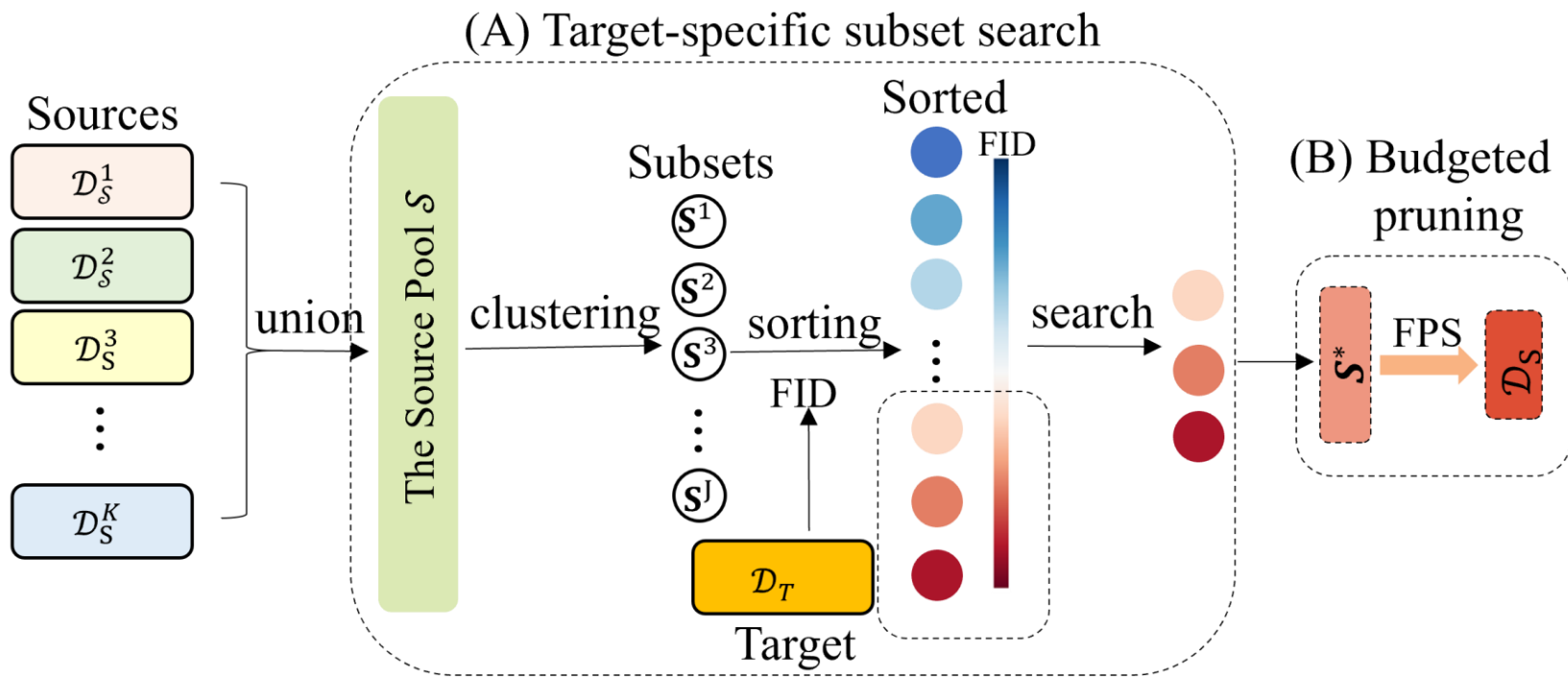Negative correlation to the domain gap (FID)

Positive correlation to the dataset size

(C)

(D)

Jointly achieve best accuracy
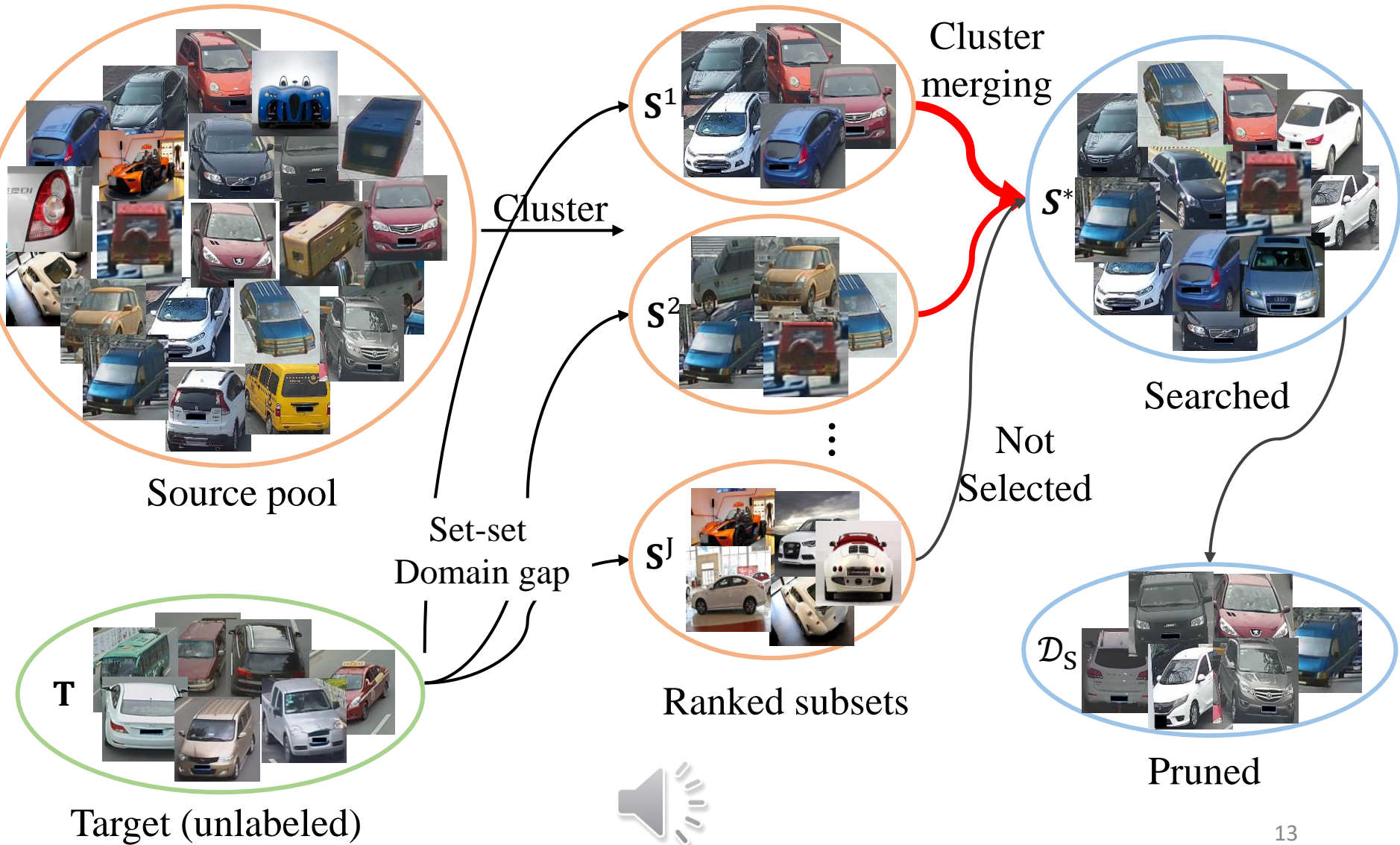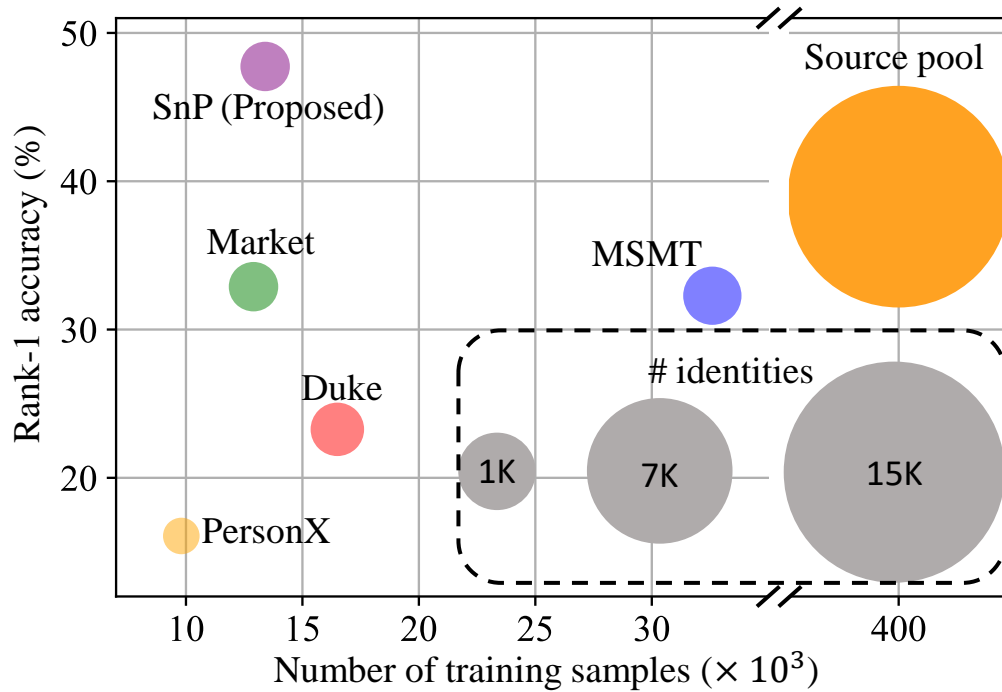
Every point is a dataset

Method: Search and Pruning (SnP)

Simple Pipeline: Search for target-specific subset; then Pruning for efficient training

Labeled data from lots of existing datasets

Source pool

Target (unlabeled)

Cluster

Set-set Domain gap

$\mathbf{S}^1$

$\mathbf{S}^2$

$\mathbf{S}^J$

Ranked subsets

Cluster merging

Not Selected

$\boldsymbol{S}^*$

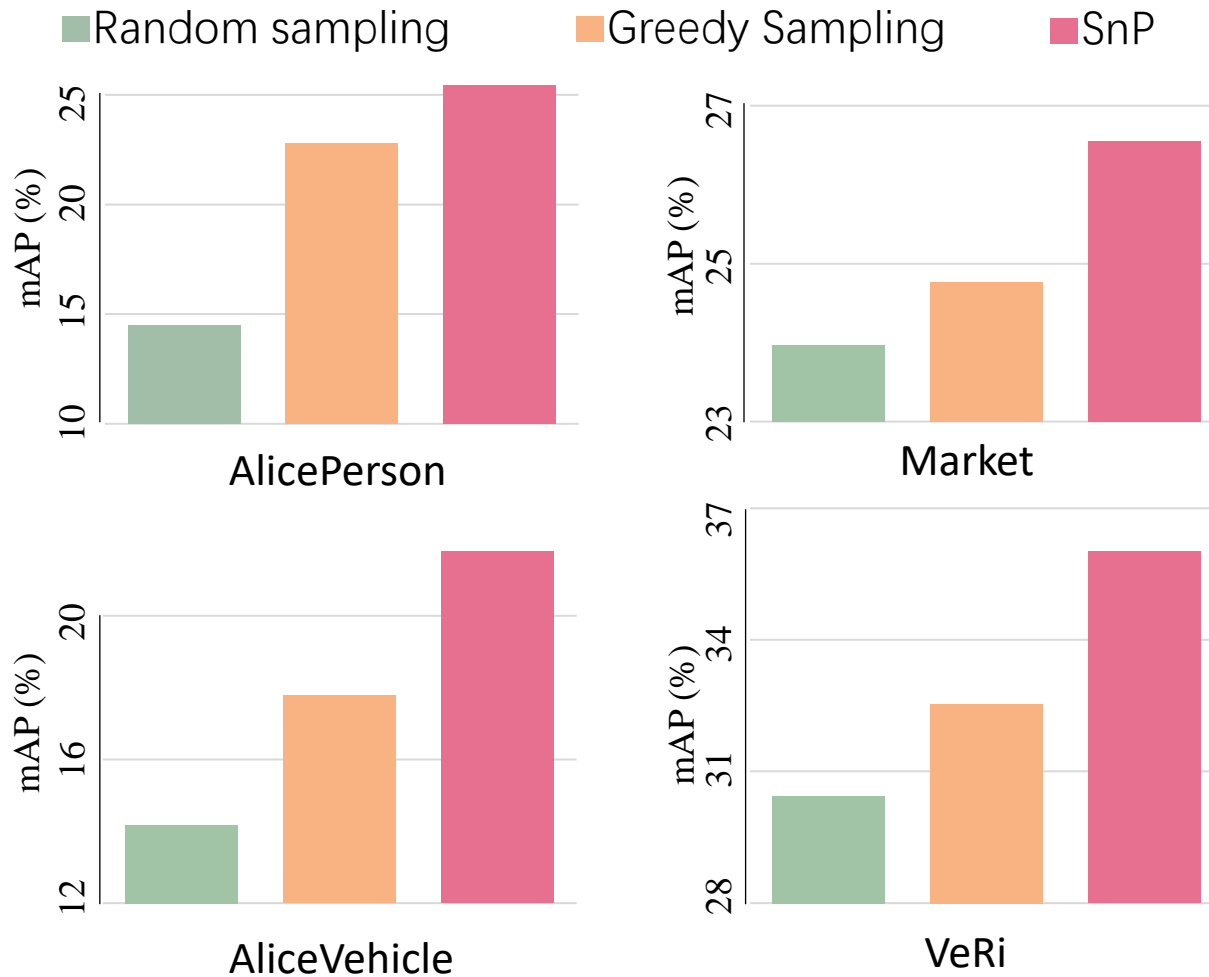Searched

$\mathcal{D}_S$

Pruned

13

Experiment

SnP has

- higher accuracy than source pool

- higher accuracy than each individual dataset

- much more small scale in the number of samples and IDs

Experiment

SnP result in better training set than both greedy sampling and random sampling

Conclusions and insights

- We study an interesting and unsolved problem:
    Training set search for unlabeled target

- We use a very simple method:
    Search and pruning framework

- Potential applications:
    Object detection, semantic segmentation, *etc*.

Scan for paper and code   Paper   Code