Unsupervised Model Evaluation

Weijian Deng

Australian National University

Australian National University

Pillars in Machine Learning

I. training

II. testing

Pillars in Machine Learning: Training

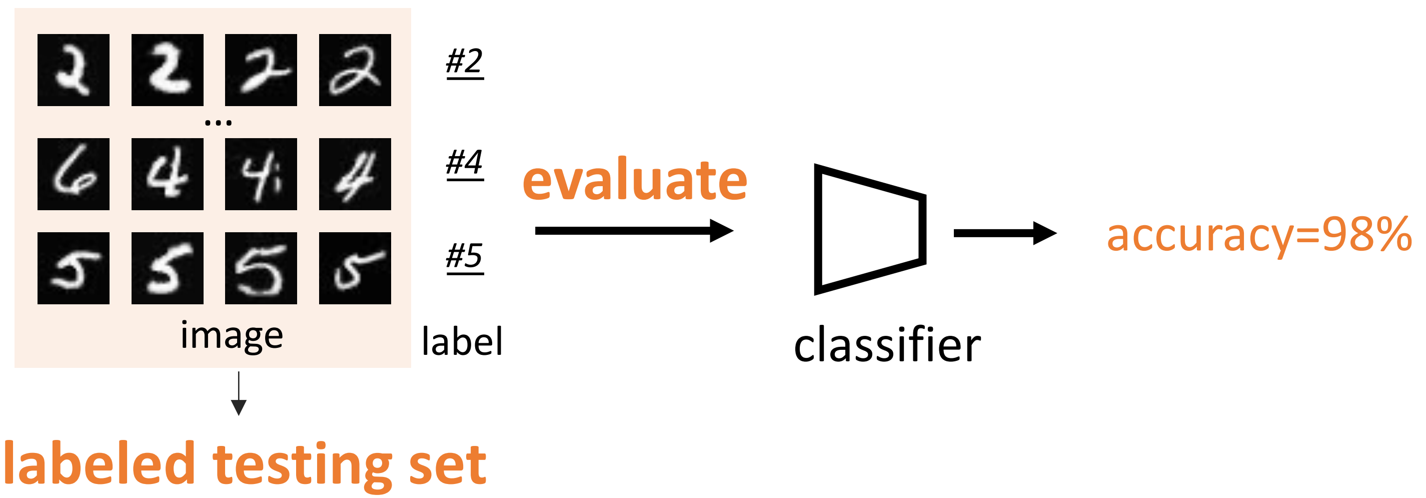
I. training

II. testing

Pillars in Machine Learning: Testing

I. training

II. testing



Supervised Evaluation

Test set is fully annotated

Ground truths are provided

image

#2

#4

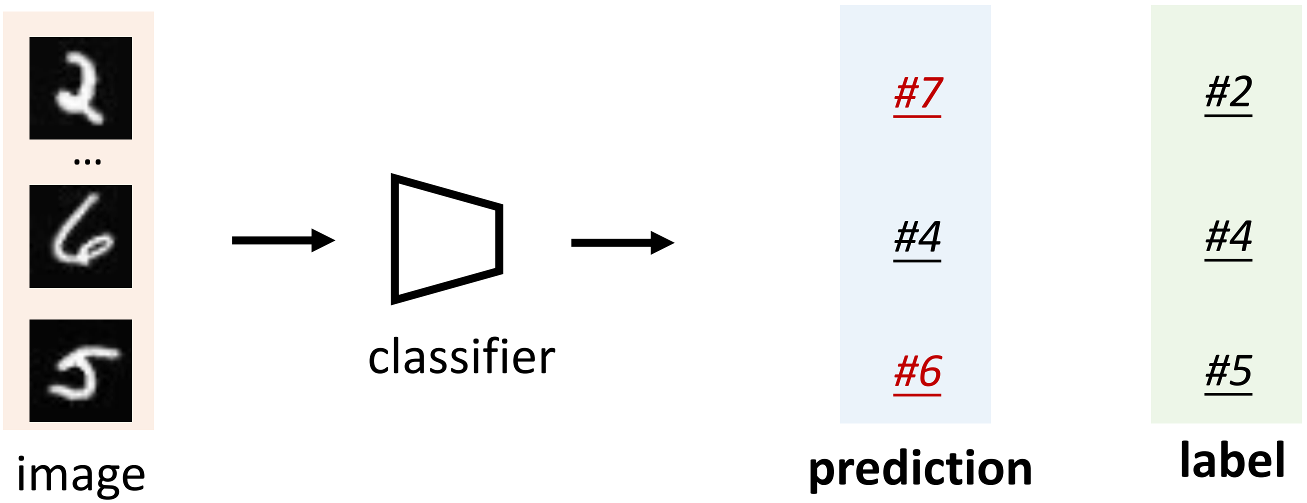
#5

label

Supervised Evaluation

Test set is fully annotated

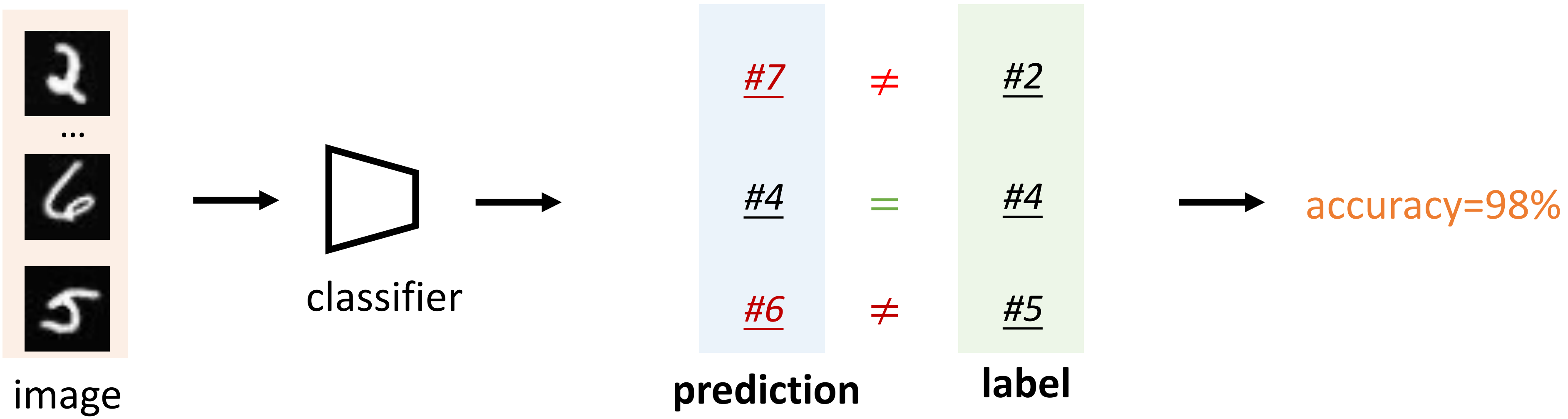
Ground truths are provided



Supervised Evaluation

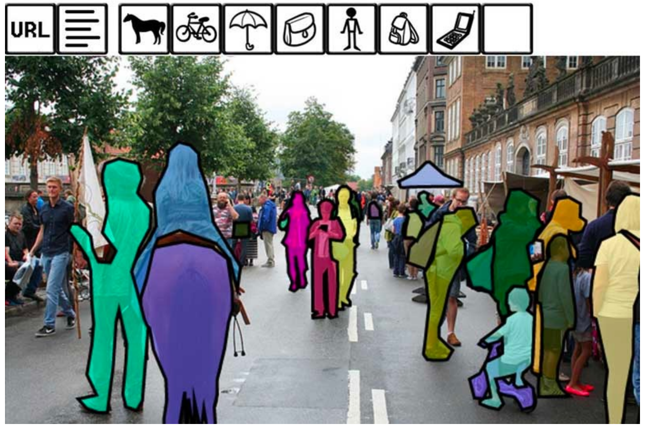
Test set is fully annotated

Ground truths are provided



In-distribution Benchmarks

ImageNet



MSCOCO

Cityscape

PASCAL

Our Research: Unsupervised Evaluation

Test set is unlabeled

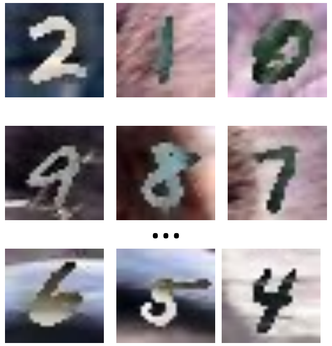
Only images are provided

→

How to evaluate model

without labels?

Unlabeled Test set 1



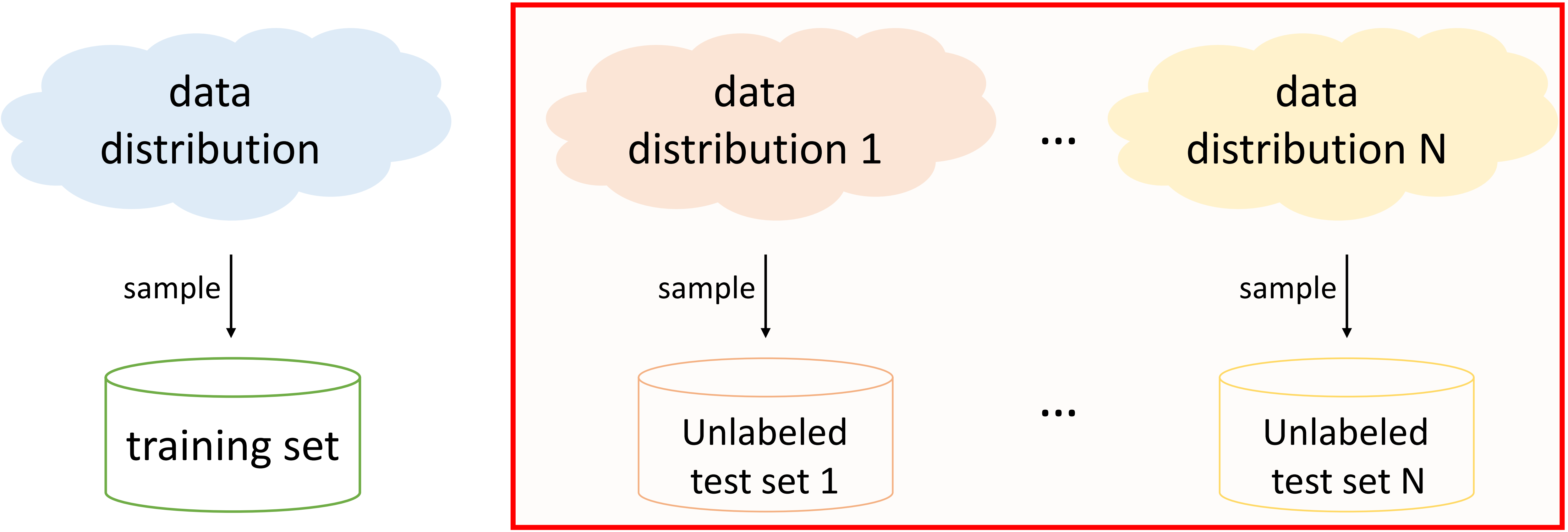
Unlabeled Test set 2

Unlabeled Test set 3

...

Unlabeled Test set 3

Evaluation Beyond Textbook



~~i.i.d. assumption~~

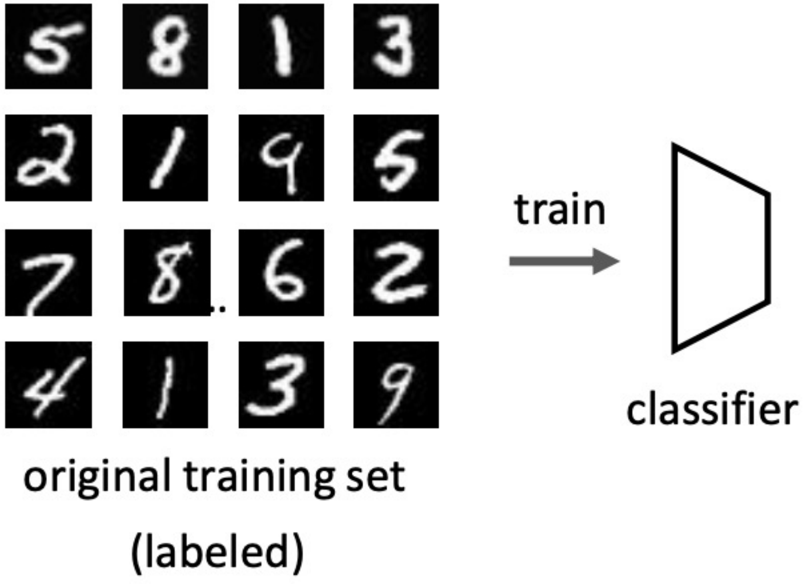
We Encounter This Problem Many Times

- Deploy face recognition model in a new airport
- Deploy a 3D object detection system to another city
- …

We can’t quantitatively measure the model accuracy like we usually do!

We need to **annotate** the test data
When the testing environment is changed, we need to **annotate again**

Our Research: Unsupervised Evaluation

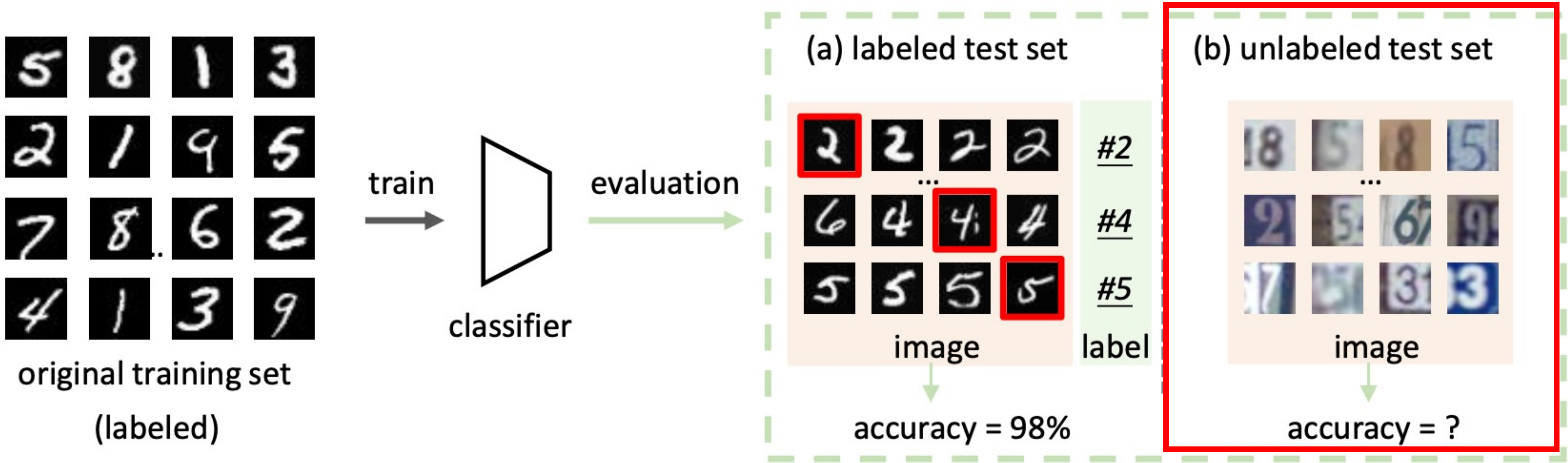


Given

- A training dataset
- A classifier trained on this dataset
- A test set without labels

Deng, Weijian, and Liang Zheng. "Are Labels Necessary for Classifier Accuracy Evaluation?", In CVPR, 2021; TPAMI 2022

Our Research: Unsupervised Evaluation



Given

- A training dataset
- A classifier trained on this dataset
- A test set without labels

We want to *estimate*:

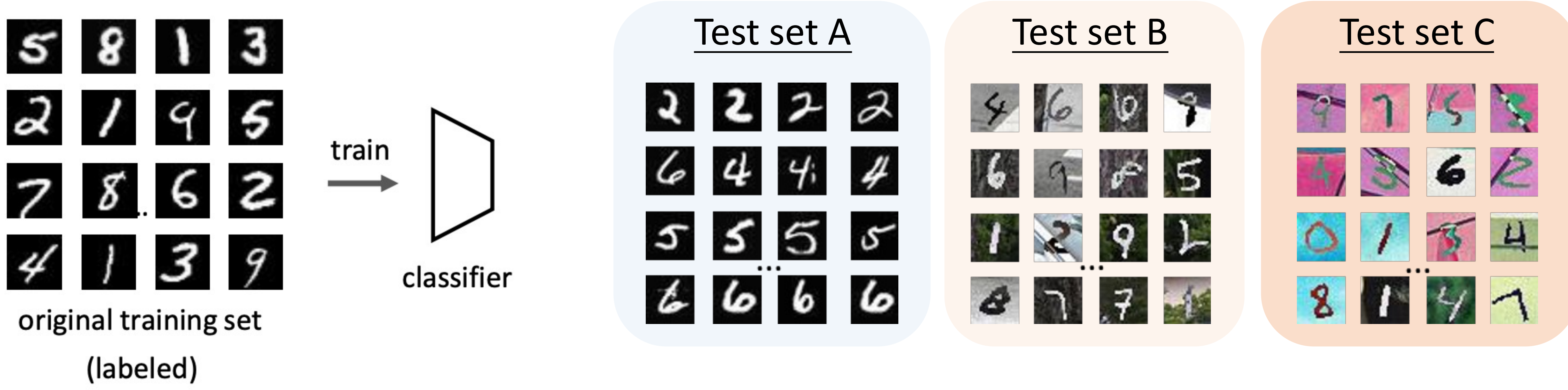
accuracy on the unlabelled test set

Deng, Weijian, and Liang Zheng. “Are Labels Necessary for Classifier Accuracy Evaluation?”, In CVPR, 2021; TPAMI 2022

Our Research: Unsupervised Evaluation

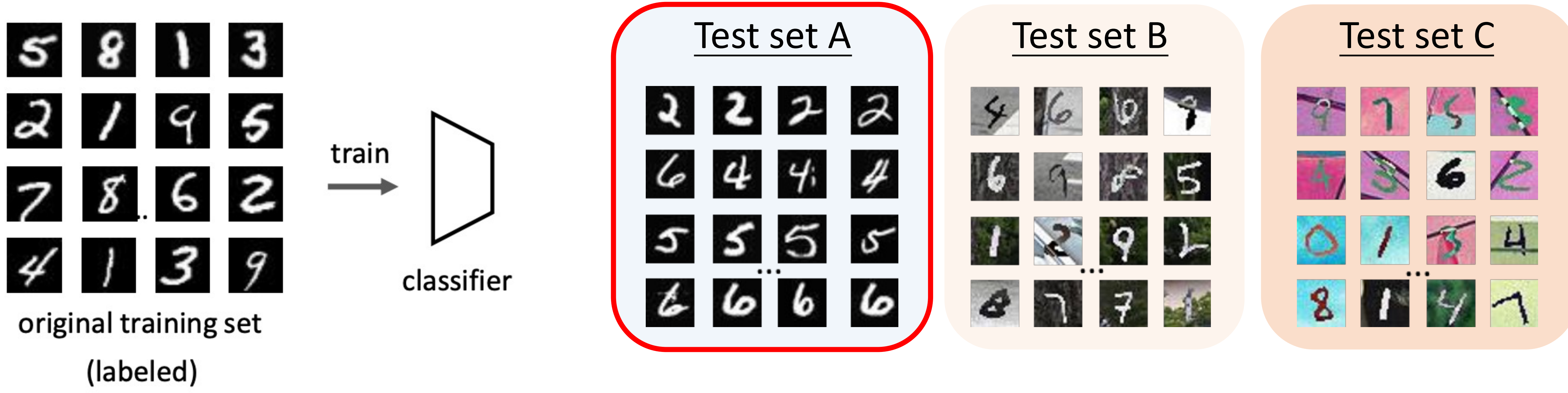
- Accuracy prediction based on dataset shift
- Self-supervision for unsupervised evaluation

Accuracy Prediction Based on Dataset Shift



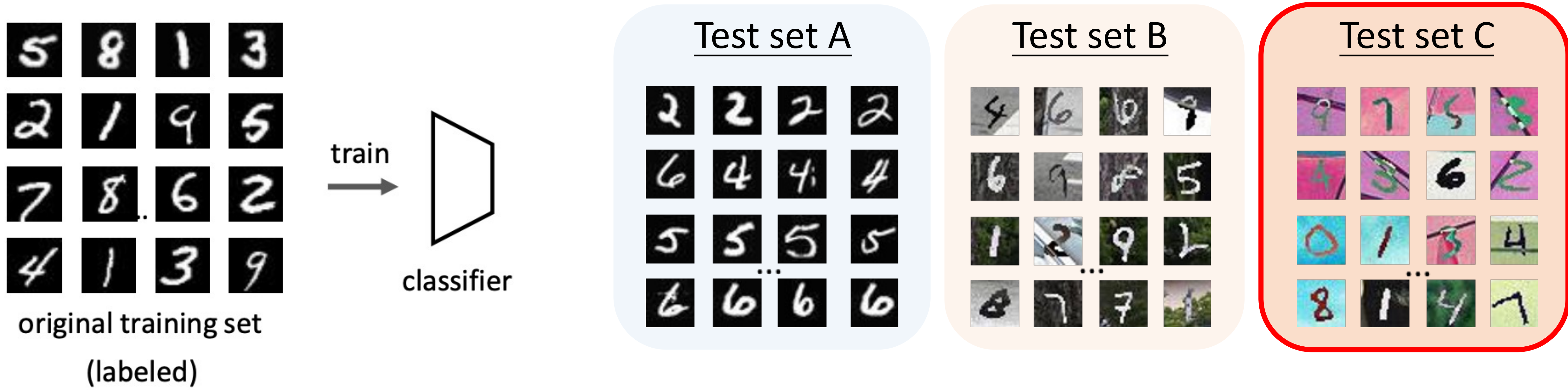
Q: Classifier performs best on…?

Accuracy Prediction Based on Dataset Shift



Test set A is more similar to training set

Accuracy Prediction Based on Dataset Shift



Test set C looks quite different from training set

Correlation Study

1. We collect **many test sets from different distributions**

2. For each test set, we obtain

   a) **its distance** with training set (*Fréchet distance*)

   b) **classification** accuracy

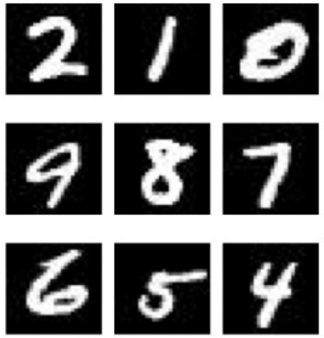
3. **Measure the accuracy relationship** between the two statistics

Correlation Study: How Can We Have Many Datasets?

- Using image transformations

original set

original set

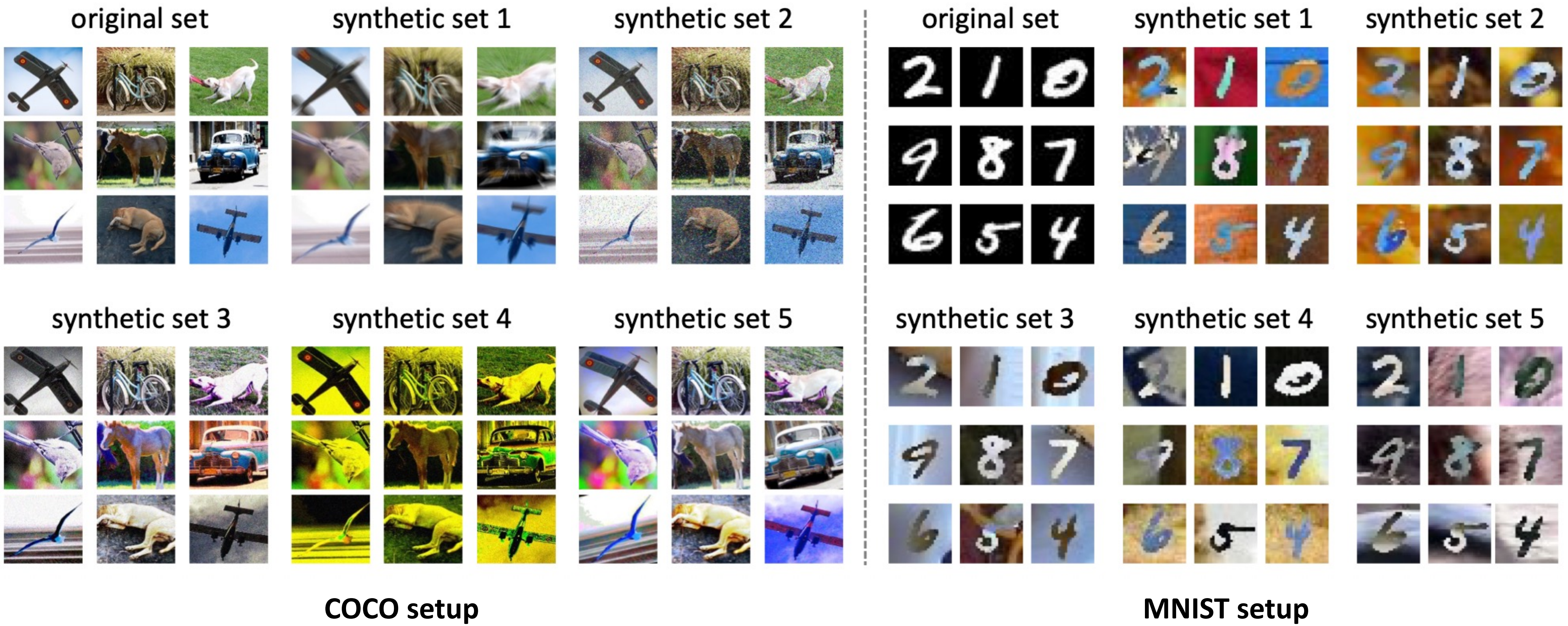


COCO setup

MNIST setup

Correlation Study: How Can We Have Many Datasets?

- Using image transformations

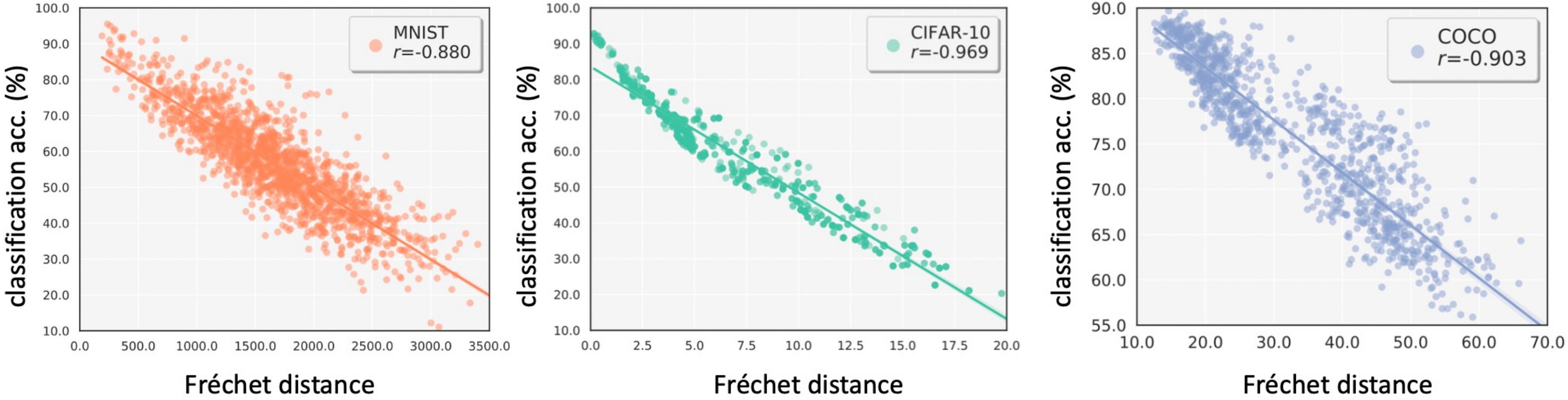


Correlation Study: How To Obtain Accuracy?

Labels of the **synthetic sets** are inherited from the **original set**

Correlation Study on Three Setups

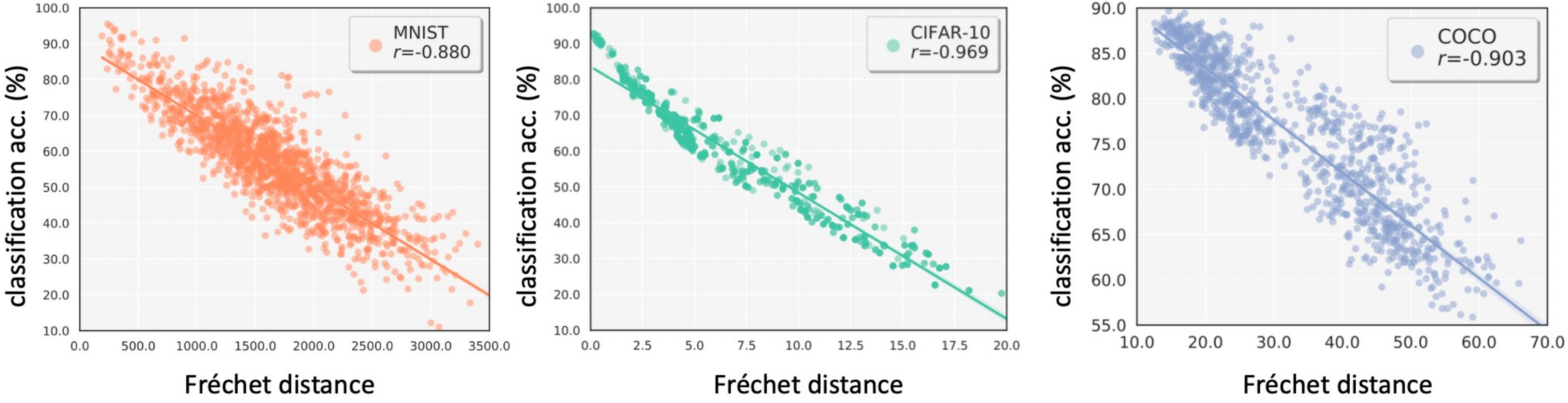
Every point is a dataset



we consistently observe a **strong negative linear relationship** (***Pearson Correlation r <0.88***) between the accuracy of two tasks

Correlation Study on Three Setups

Every point is a dataset



This indicates that the classifier tends to gain a **high accuracy** on the sample set which has a **low distribution shift** with training set.

Accuracy Estimation on Unseen Test Sets

- **Linear regression**
- **Network regression**

Accuracy Estimation on Unseen Test Sets

- **Linear regression**

  Fréchet distance (FD) between the test set and the original training set

$$a_{linear} = A_{linear}(\boldsymbol{f}) = w_1 \boxed{f_{linear}} + w_0$$

Fréchet distance

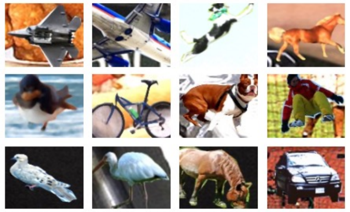
$$f_{linear} = \mathrm{FD}(\mathcal{D}_{ori}, \mathcal{D}) = \|\boldsymbol{\mu}_{ori} - \boldsymbol{\mu}\|_2^2 + Tr(\boldsymbol{\Sigma}_{ori} + \boldsymbol{\Sigma} - 2(\boldsymbol{\Sigma}_{ori}\boldsymbol{\Sigma}))^{\frac{1}{2}}$$

Accuracy Estimation on Unseen Test Sets

- Linear regression
- **Network regression**

FD + mean + covariance (sum) for representing each dataset

We calculate $\boldsymbol{\sigma}$ by taking a weighted summation of each row of $\boldsymbol{\Sigma}$ to produce a single vector



$$\longrightarrow \boldsymbol{f}_{neural} = [f_{linear}; \boldsymbol{\mu}; \boldsymbol{\sigma}]$$

- We use neural network regression

$$a_{neural} = A_{neural}(\boldsymbol{f}_{neural})$$

Accuracy Estimation on Unseen Test Sets

- Linear regression achieves promising estimations

Training set	Seed set	Test sets
COCO training set	COCO validation set	PASCAL, ImageNet, and Caltech

Image transformations

Many synthesized test sets

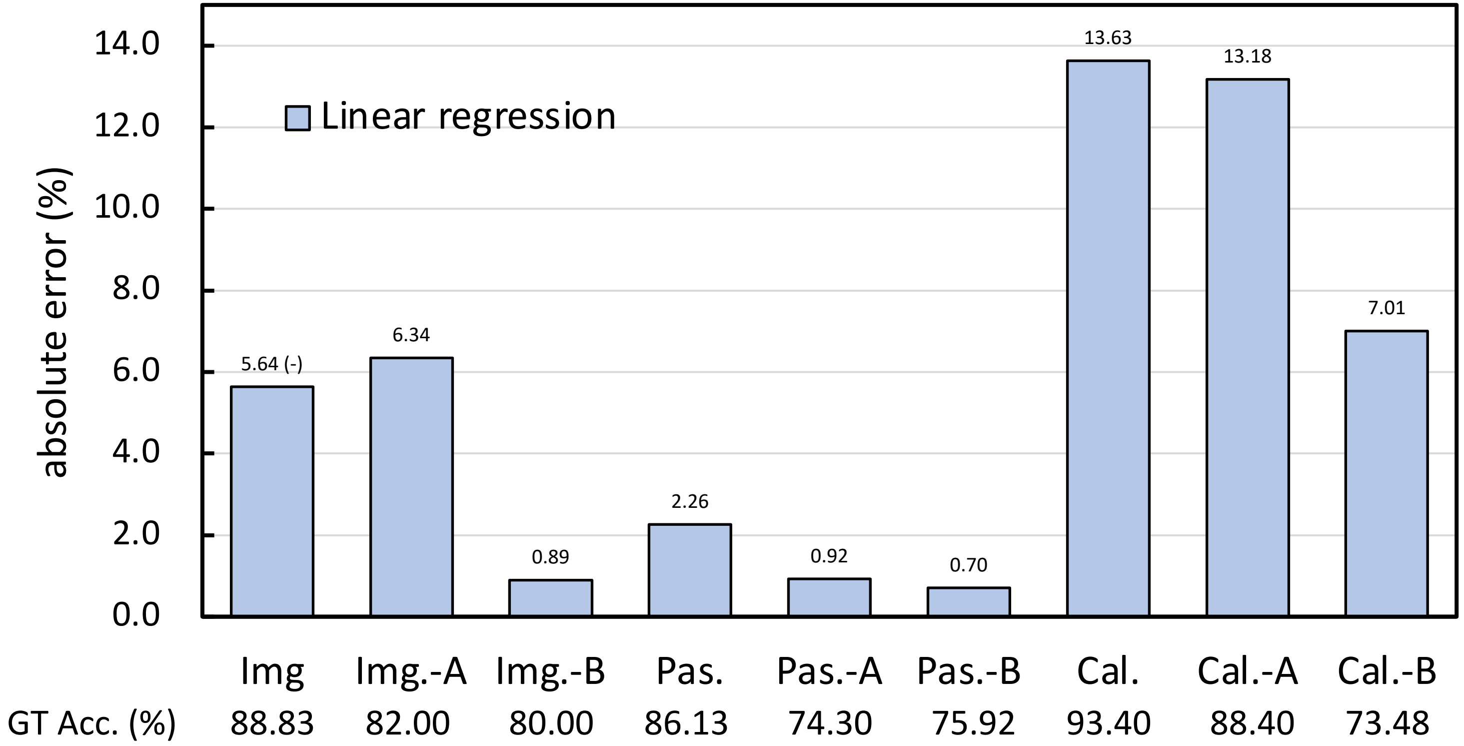
train

Regression models

Accuracy Estimation on Unseen Test Sets

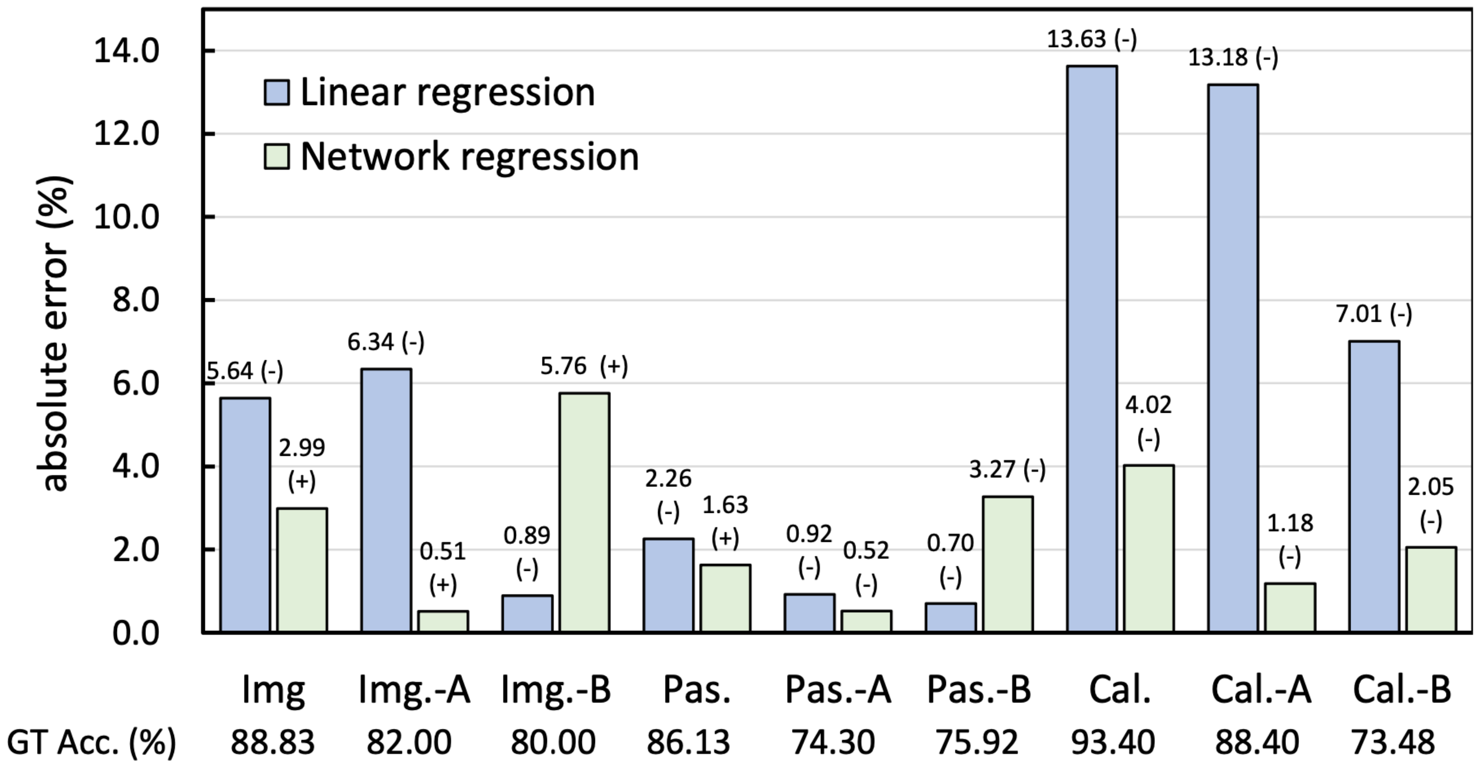
- Linear regression achieves promising estimations

Training set	Seed set	Test sets
COCO training set	COCO validation set	PASCAL, ImageNet, and Caltech



Accuracy Estimation on Unseen Test Sets

- Linear regression achieves promising estimations
- **Network regression makes more accurate predictions**

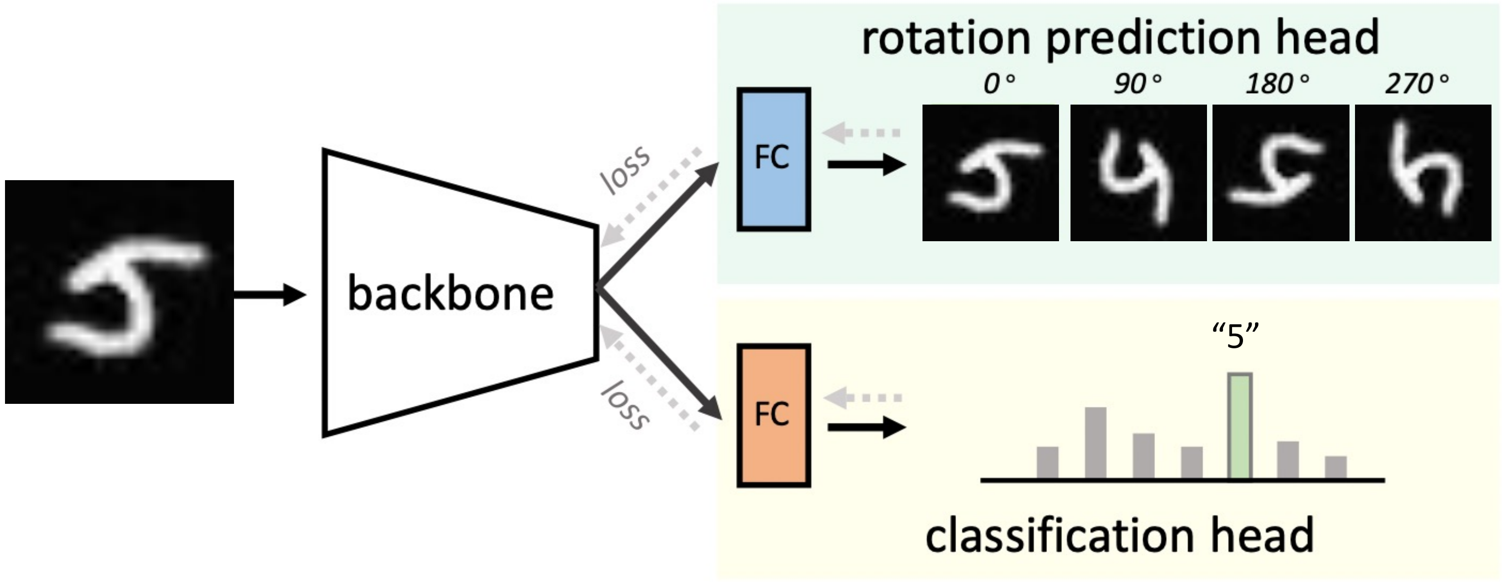


Our Research: Unsupervised Evaluation

- Accuracy prediction based on dataset shift
- **Self-supervision for unsupervised evaluation**

Self-Supervision for Unsupervised Classifier Evaluation

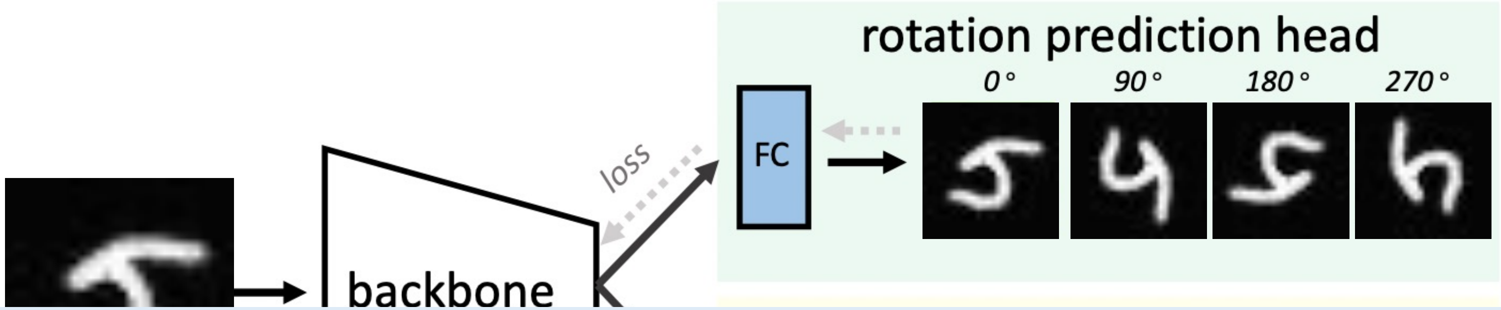
- **Multi-task network structure**



Deng, Weijian, Stephen Gould, and Liang Zheng. "What Does Rotation Prediction Tell Us about Classifier Accuracy under Varying Testing Environments?." ICML, 2021.

Self-Supervision for Unsupervised Classifier Evaluation

- **Multi-task network structure**



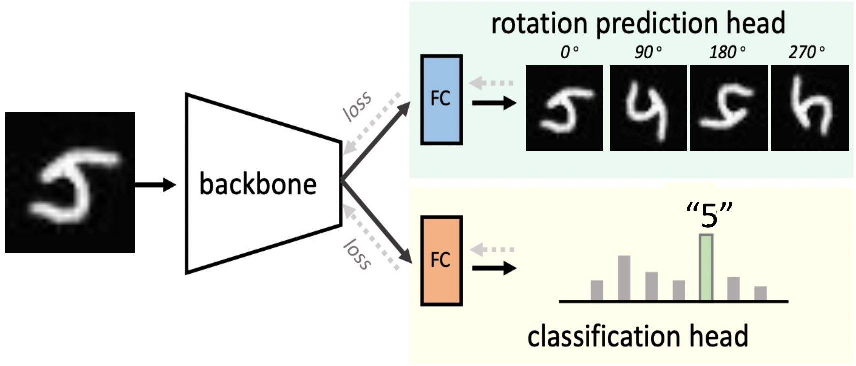
Rotation prediction is self-supervised:

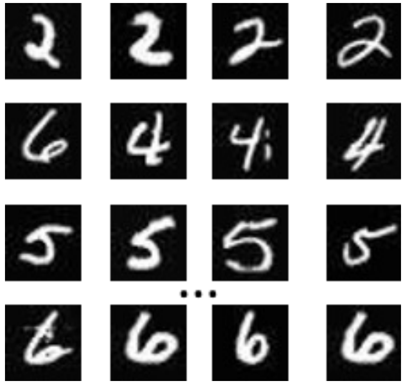
we can *obtain its rotation labels freely* and

calculate its *accuracy on any test set*

Deng, Weijian, Stephen Gould, and Liang Zheng. "What Does Rotation Prediction Tell Us about Classifier Accuracy under Varying Testing Environments?." ICML, 2021.

Motivation



	Test set 1	Test set 2	Test set 3
			
rotation prediction accuracy	95%	85%	75%
recognition accuracy:	90%	80%	70%

Correlation Study

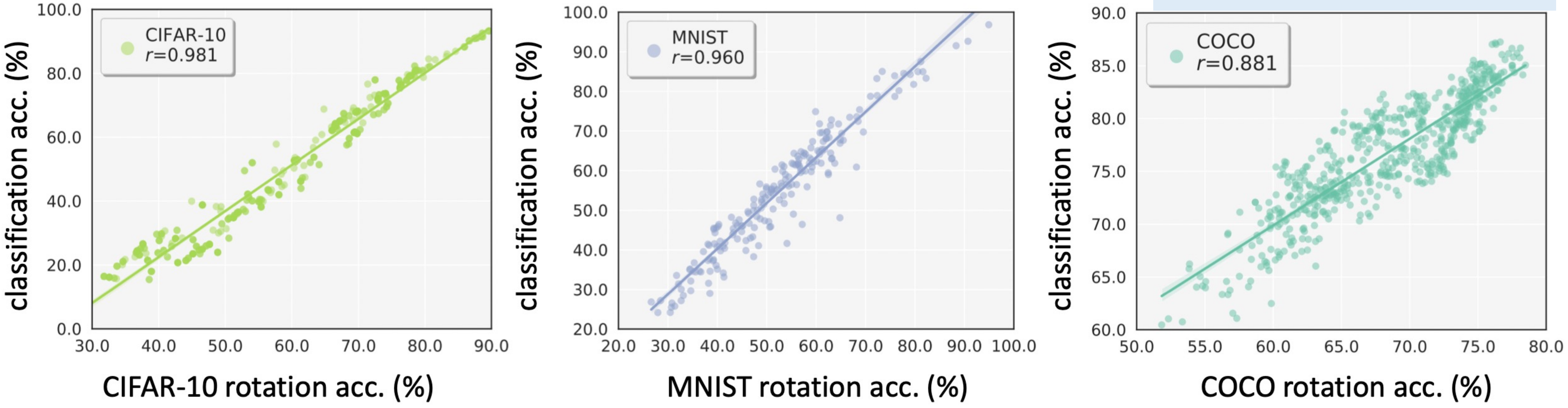
1. We collect **many test sets from different distributions**

2. Test our multi-task network on them and obtain
   a) sematic classification accuracy
   b) rotation prediction accuracy

3. **Measure the accuracy relationship** between two types of tasks

Correlation Study on Three Setups

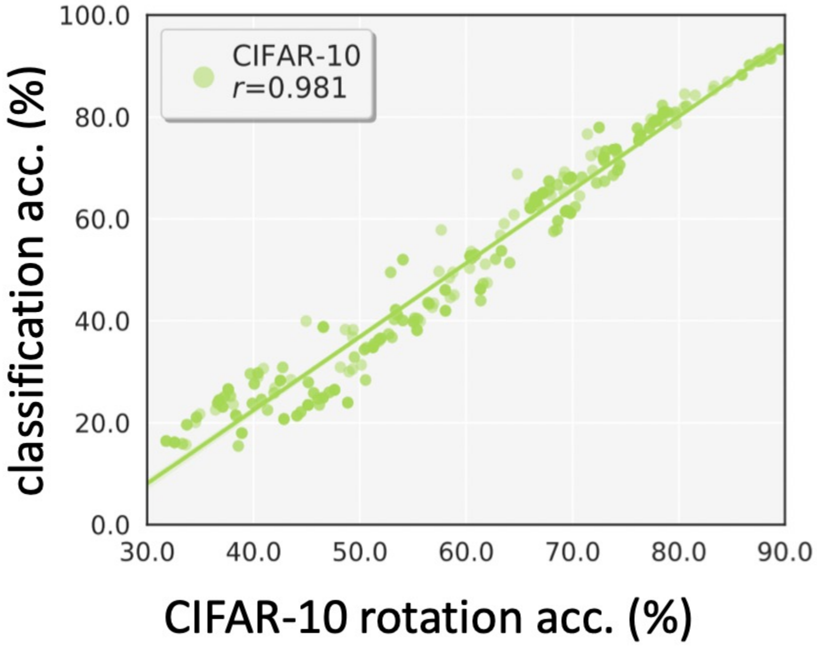
Every point is a dataset

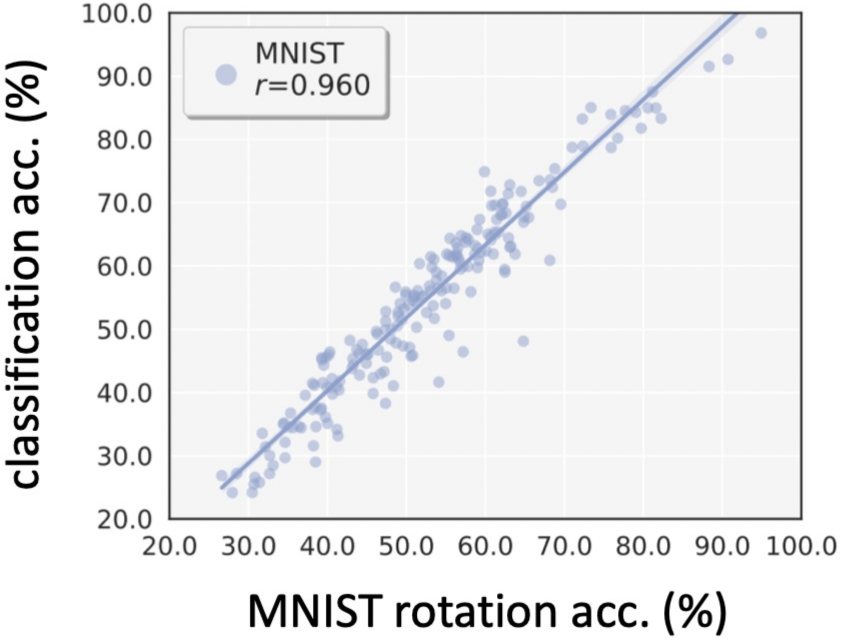


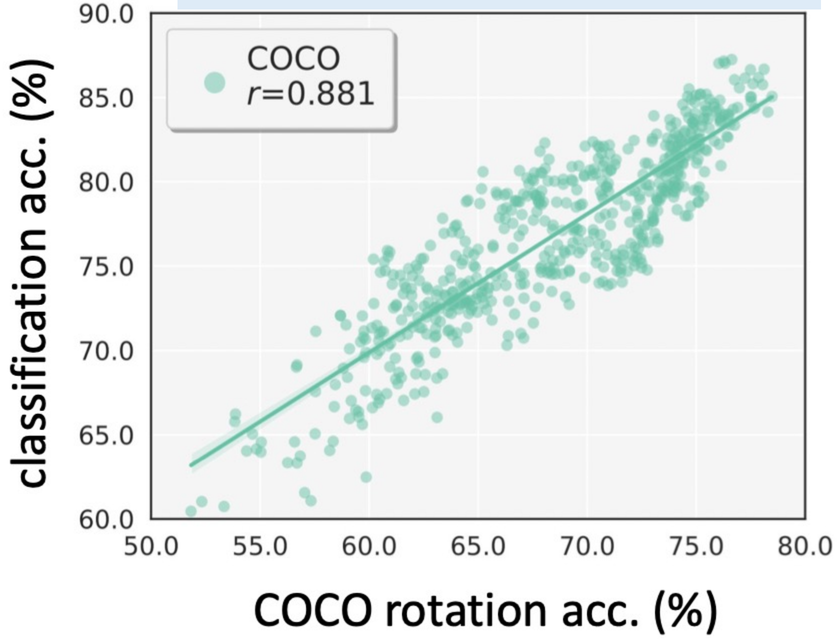
we consistently observe a **strong linear relationship** (***Pearson Correlation r > 0.88***) between the accuracy of two tasks

Correlation Study on Three Setups

Every point is a dataset







If the multi-task **network is good at predicting rotations,** it is most likely to **achieve good object recognition accuracy** under the same environment, and vice versa

Our Solution for Accuracy Estimation: Linear Regression

- **Method:**

Predict classifier performance from rotation prediction accuracy

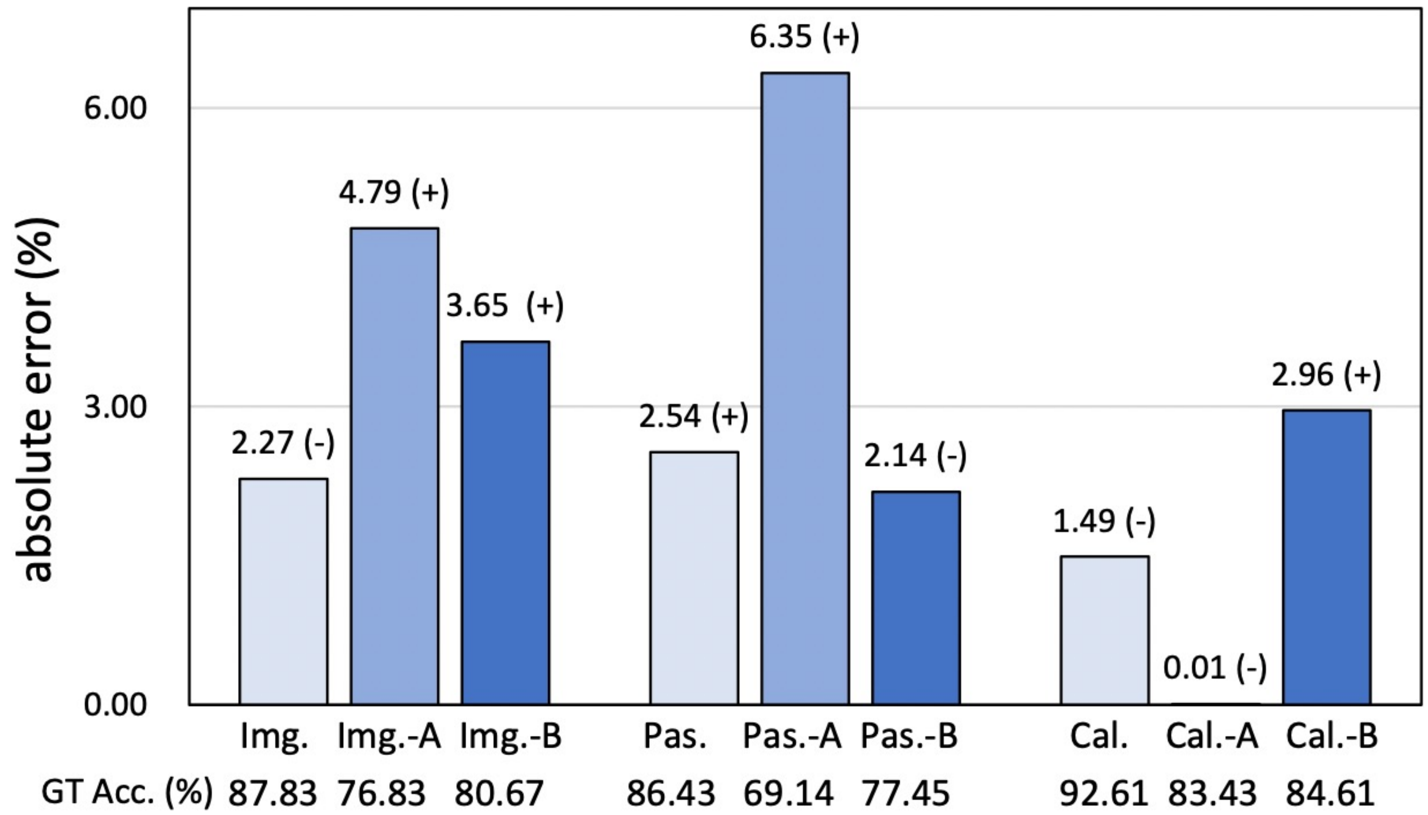
We thus can use linear regression to predict accuracy

$$\boxed{a^{cls} = w_1 a^{rot} + w_0,}$$

where $w_1, w_0 \in \mathbb{R}$ are linear regression parameters

Accuracy Estimation on Unseen Test Sets

- Linear regression achieves promising estimations



Conclusions and Insights

- We study a very interesting problem:

  Evaluating model performance *without* ground truths

- We introduce a very simple method:

  Dataset-level regression (Linear regression and Neural network regression)

- Potential Applications:

  Other tasks: object retrieval, detection, segmentation, etc.

Thank you!

The code is available at
https://weijiandeng.xyz