Australian National University

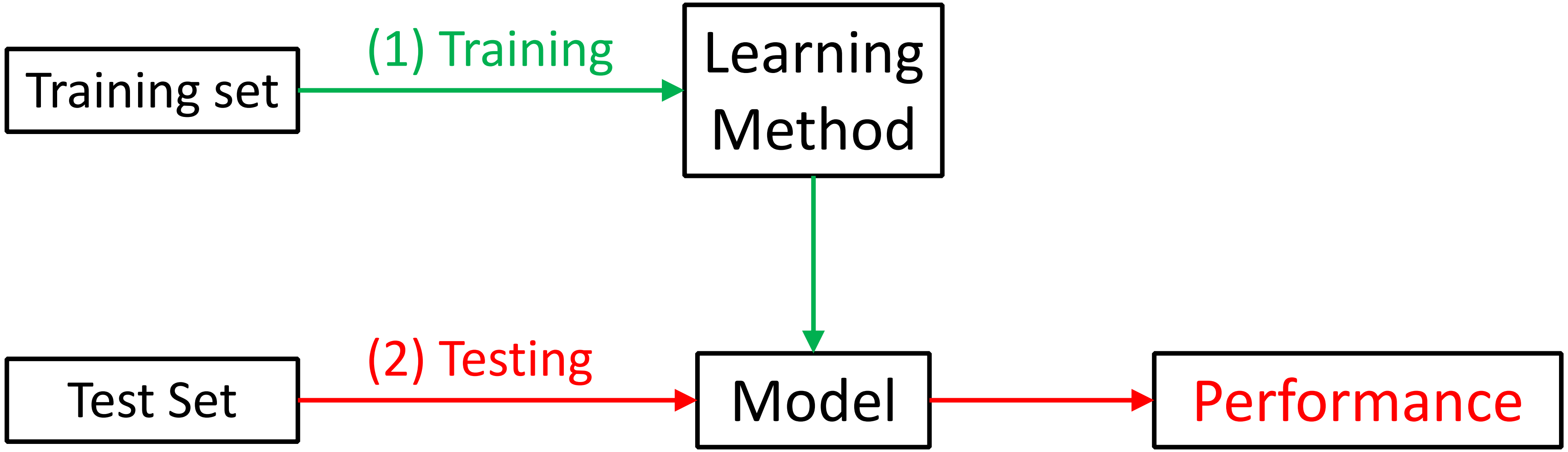
ICML International Conference On Machine Learning

What Does Rotation Prediction Tell Us about Classifier Accuracy under Varying Testing Environments?

Weijian Deng Stephen Gould Liang Zheng

Australian National University

Pillars in machine learning

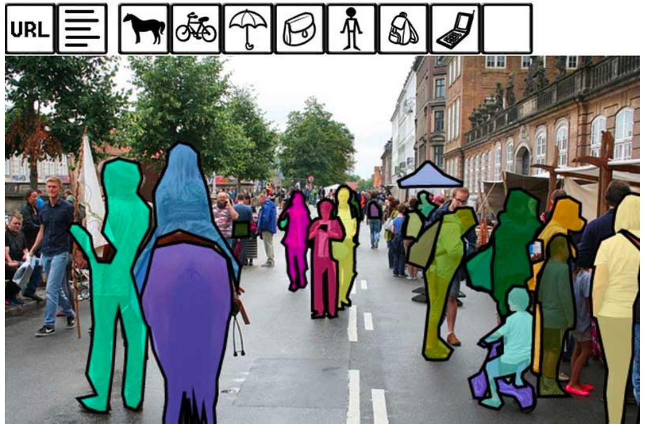


Is evaluation feasible?

- Yes

Labelled test set → *Ground truths are provided*

ImageNet



MSCOCO

Is evaluation feasible?

- No

Unlabeled test images → *Ground truths are not provided*

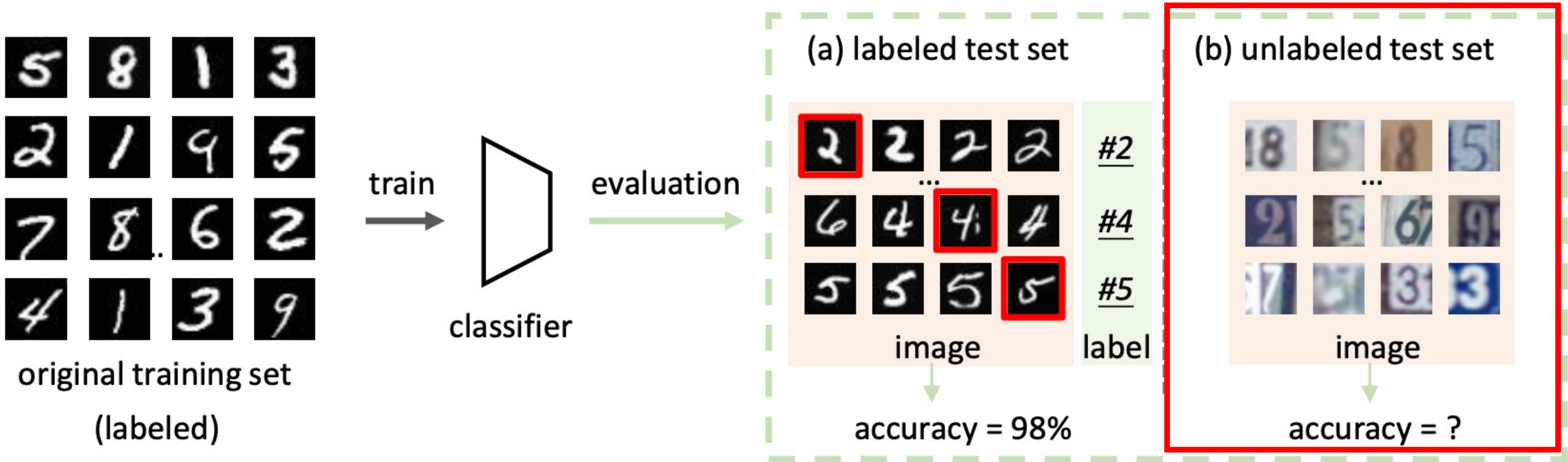
We encounter this problem many times

- Deploy face recognition model in an airport
- Deploy a 3D object detection system to a new 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**

Self-supervision for unsupervised classifier evaluation



Given

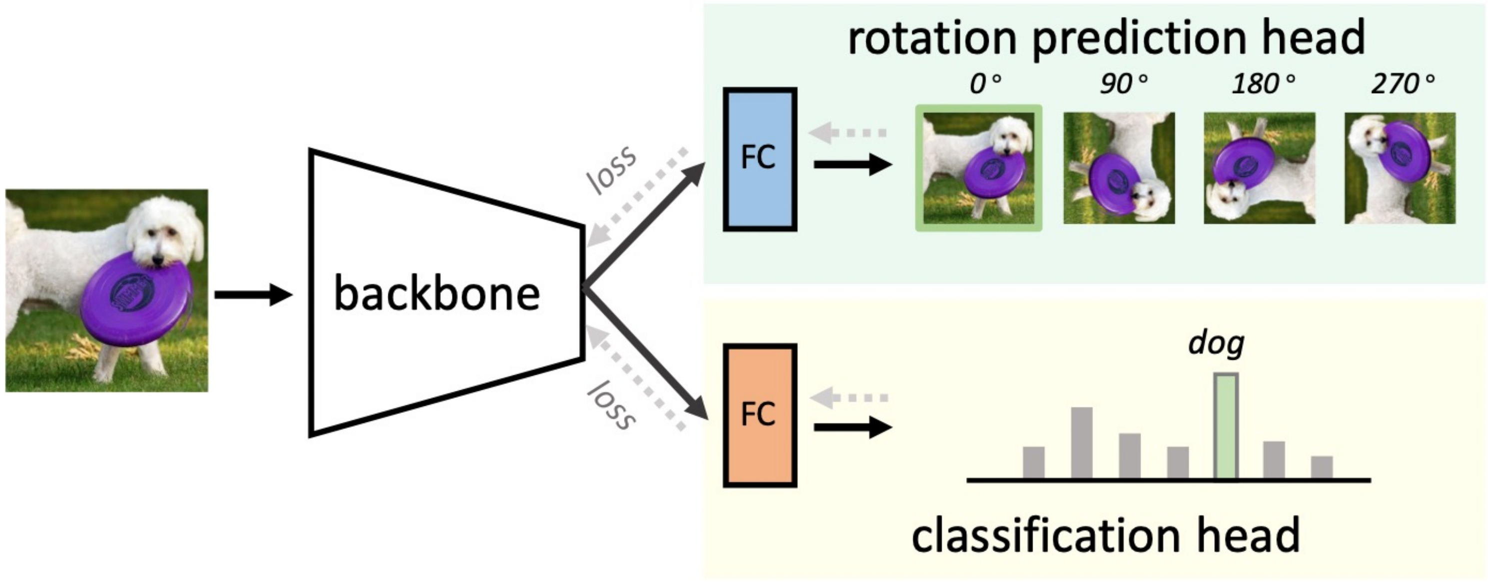
- A training dataset
- A classifier trained on this dataset
- A test set without labels

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

Self-supervision for unsupervised classifier evaluation



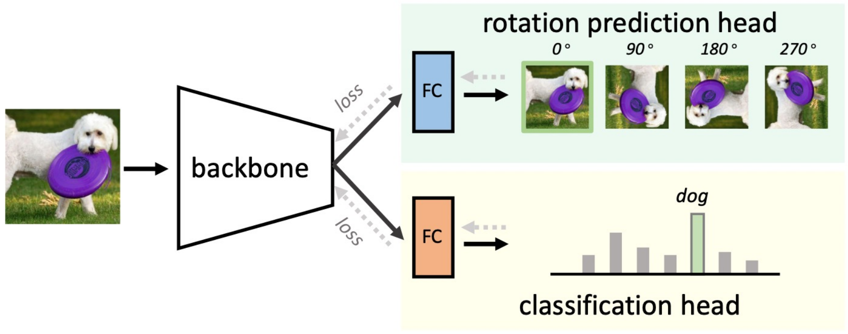
multi-task network structure

The self- supervised task *should*

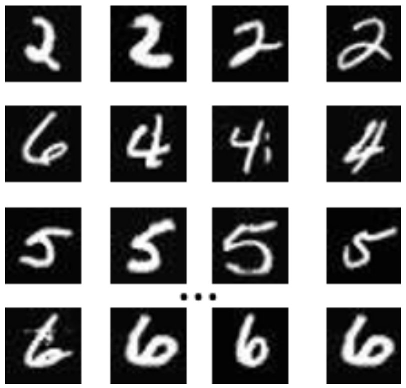
1) introduce no learning complexity for the main classification;
2) require minimal structure change;
3) not degrade classification accuracy

rotation prediction

Motivation



Test set 1



Test set 2

Test set 3

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

Motivation

Rotation prediction is self-supervised:

we can *obtain its rotation labels freely* and

calculate its *accuracy on any test set*

If rotation prediction accuracy *is correlated with*

semantic classification accuracy,

then we can **predict** the classifier performance from

the accuracy of rotation prediction

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

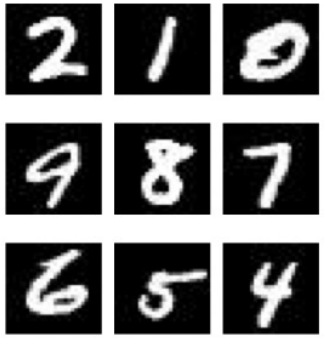
3. **Measure accuracy relationship** between two types of tasks

Correlation study: how can we have many datasets?

- Using image transformations

original set

original set

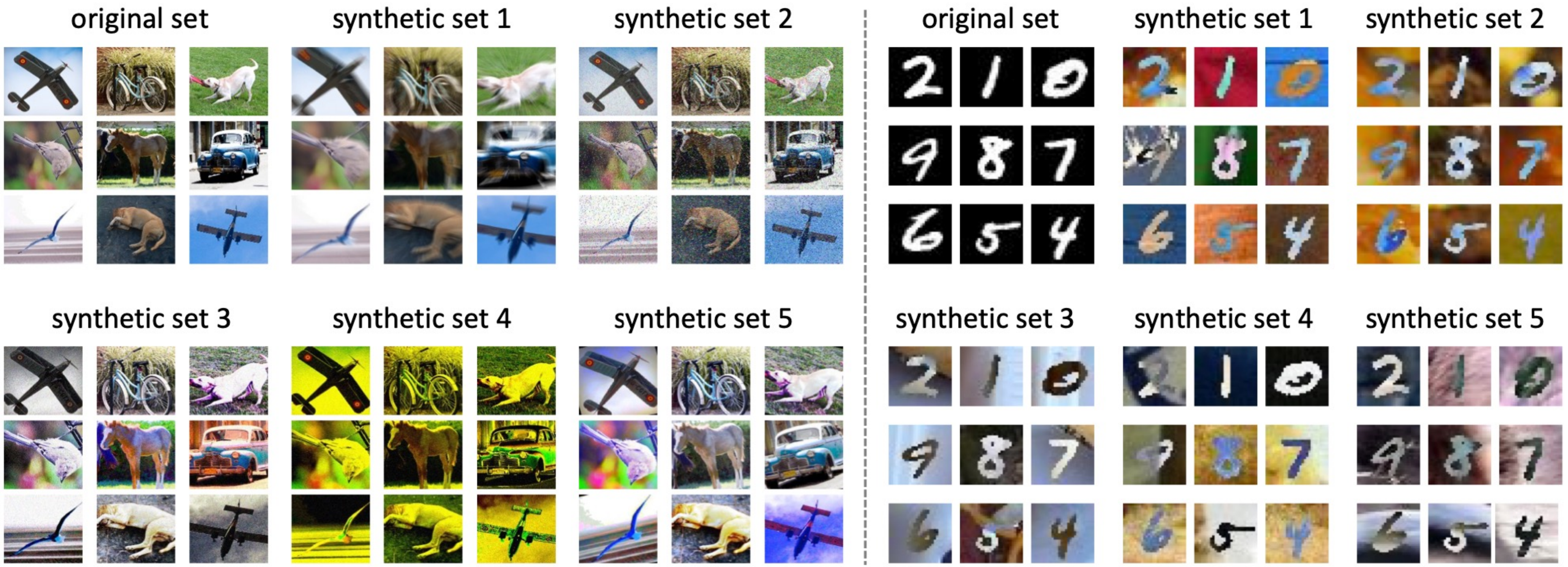


COCO setup

MNIST setup

Correlation study: how can we have many datasets?

- Using image transformations



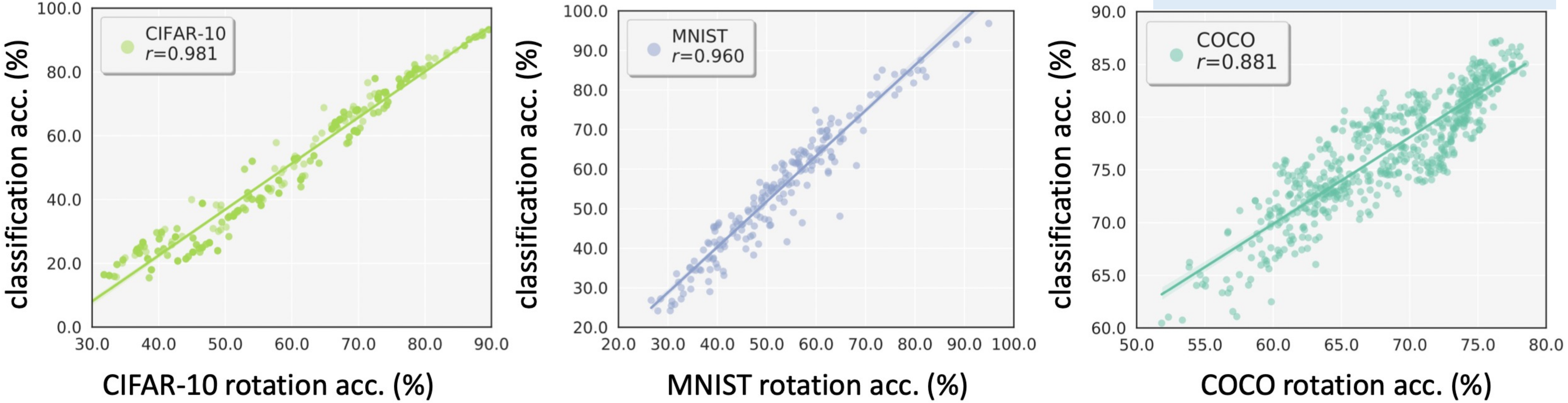
COCO setup **MNIST setup**

Correlation study: how to obtain accuracy?

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

Correlation study on three setups

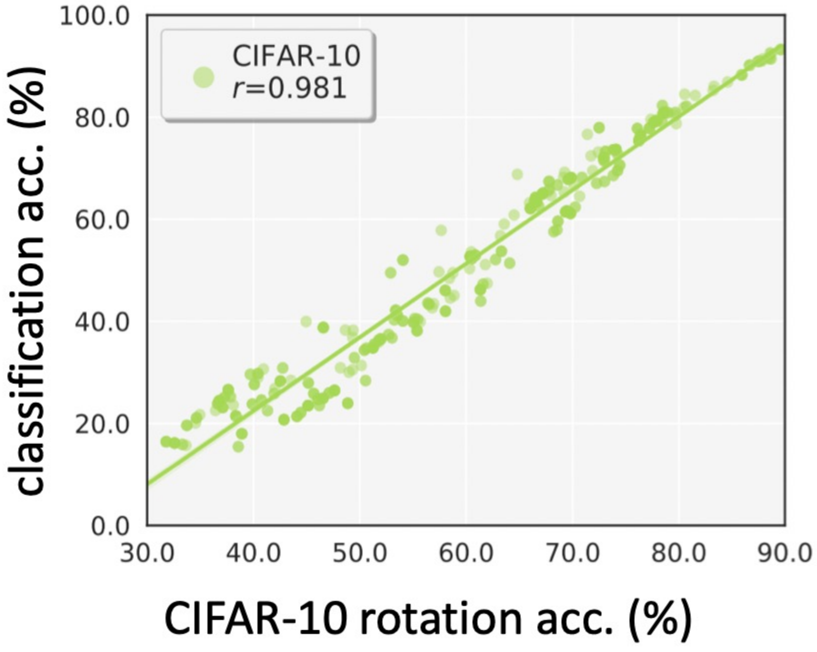
Every point is a dataset

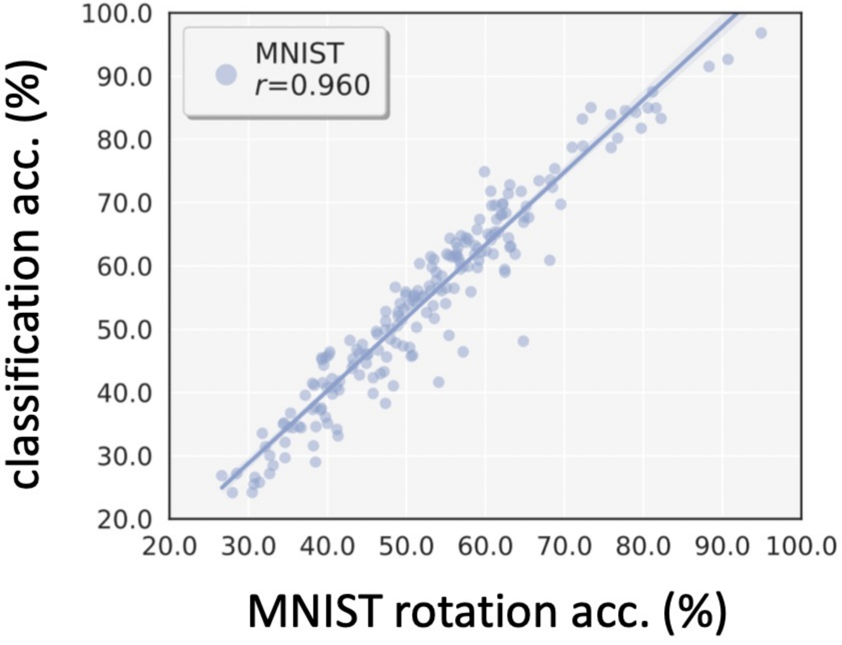


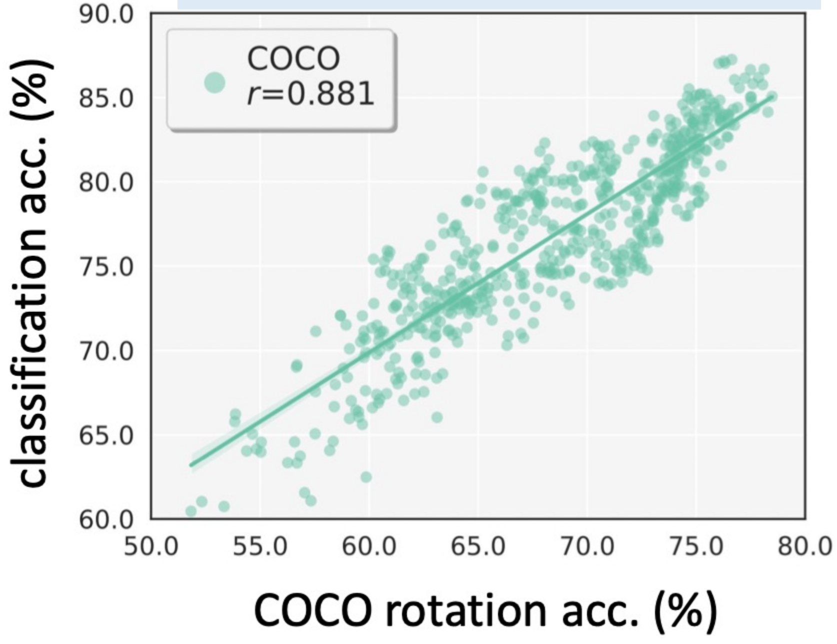
we consistently observe a **strong linear relationship** (***Pearson Correlation r > 0.88***) between the accuracy of two tasks

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

Correlation study with different backbones

CIFAR-10 Setup

	VGG11	VGG19	ResNet26	ResNet44	Dense40
Class. Acc.	92.53	92.51	92.84	93.73	94.75
Rot. Acc.	91.32	92.07	87.84	88.81	91.28
Cor. (r)	0.990	0.987	0.975	0.981	0.981

The strong linear correlation **is maintained** when using different backbones.

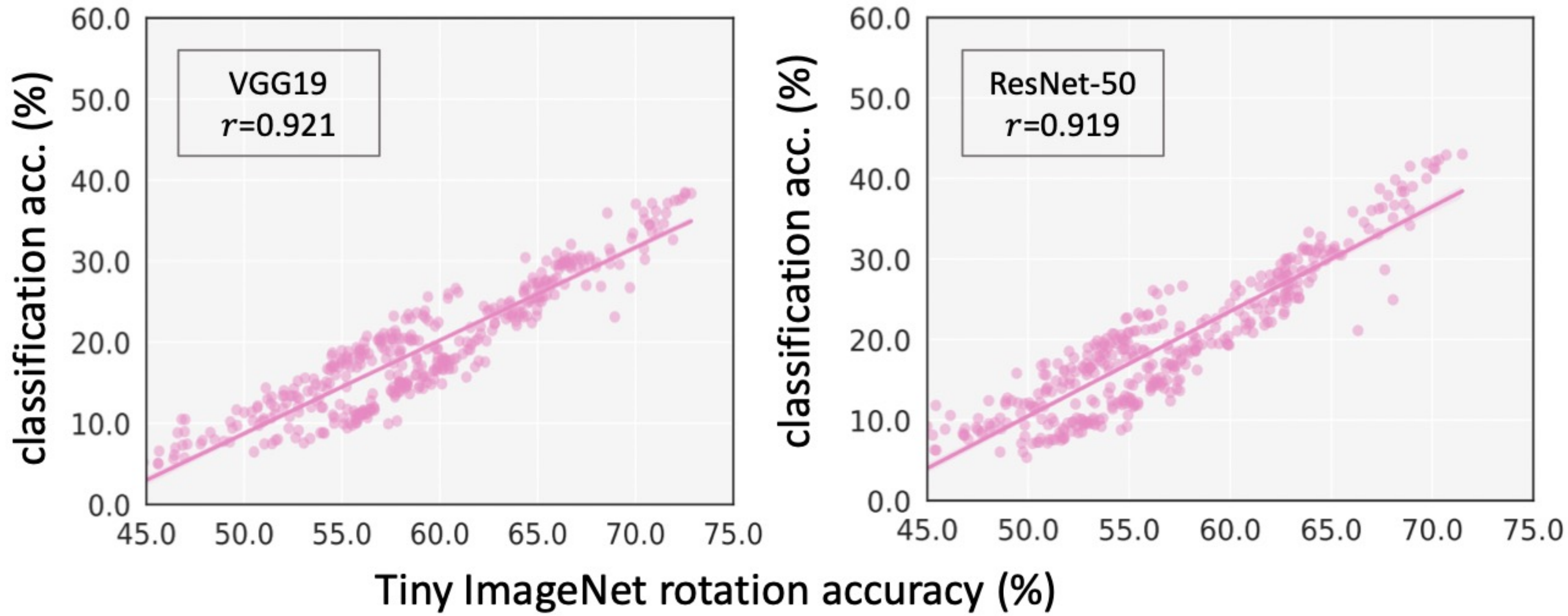
Correlation when the number of classes is large

CIFAR-100 Setup

Backbone	CIFAR-10	CIFAR-100		
	Cor. (r)	Cor. (r)	Class Acc.	Rot. Acc.
ResNet26	0.975	0.918	69.31	73.18
ResNet44	0.981	0.910	71.38	75.60
Dense40	0.981	0.950	74.55	75.20

Correlation when the number of classes is large

Tiny-ImageNet (200 classes)



When the number of categories is huge (*e.g.*, 10K (Deng et al., 2010)), the **correlation might decrease** but it will **still have a high value**.

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

Experiment on accuracy estimation

Settings	Training set	Seed set	Test sets
MNIST	MNIST training set	MNIST test set	SVHN and USPS
COCO	COCO training set	COCO validation set	PASCAL, ImageNet, and Caltech
CIFAR-10	CIFAR-10 training set	CIFAR-10 test set	CIFAR10.1 (a new test set)

We use root mean squared error (RMSE) to evaluate the accuracy prediction

Experiment on accuracy estimation

Train Set	MNIST			CIFAR-10		COCO			
Unseen Test Set	SVHN	USPS	RMSE↓	CIFAR-10.1	RMSE↓	Caltech	Pascal	ImageNet	RMSE↓
Ground-truth Accuracy	23.06	65.52	-	88.15	-	92.61	86.43	87.83	-
Prediction ($\tau_1 = 0.8$)	33.64	44.34	16.74	91.15	3.00	89.36	83.98	85.17	2.81
Prediction ($\tau_1 = 0.9$)	22.07	30.39	24.85	86.85	1.30	84.30	78.00	79.83	8.25
Entropy ($\tau_2 = 0.2$)	26.63	33.23	22.97	89.20	1.05	86.80	80.14	82.50	5.82
Entropy ($\tau_2 = 0.3$)	40.35	46.87	17.98	93.80	5.65	92.49	86.21	88.50	0.41

"Predicted Score" and "Entropy Score":
two intuitive pseudo label methods

If the **maximum value** of the softmax outputs (Predicted Score) is **greater** than τ_1, we view this sample as correctly classified.

If the **entropy value** of the softmax outputs (entropy Score) is **lower** than τ_2, we view this sample as correctly classified.

Experiment on accuracy estimation

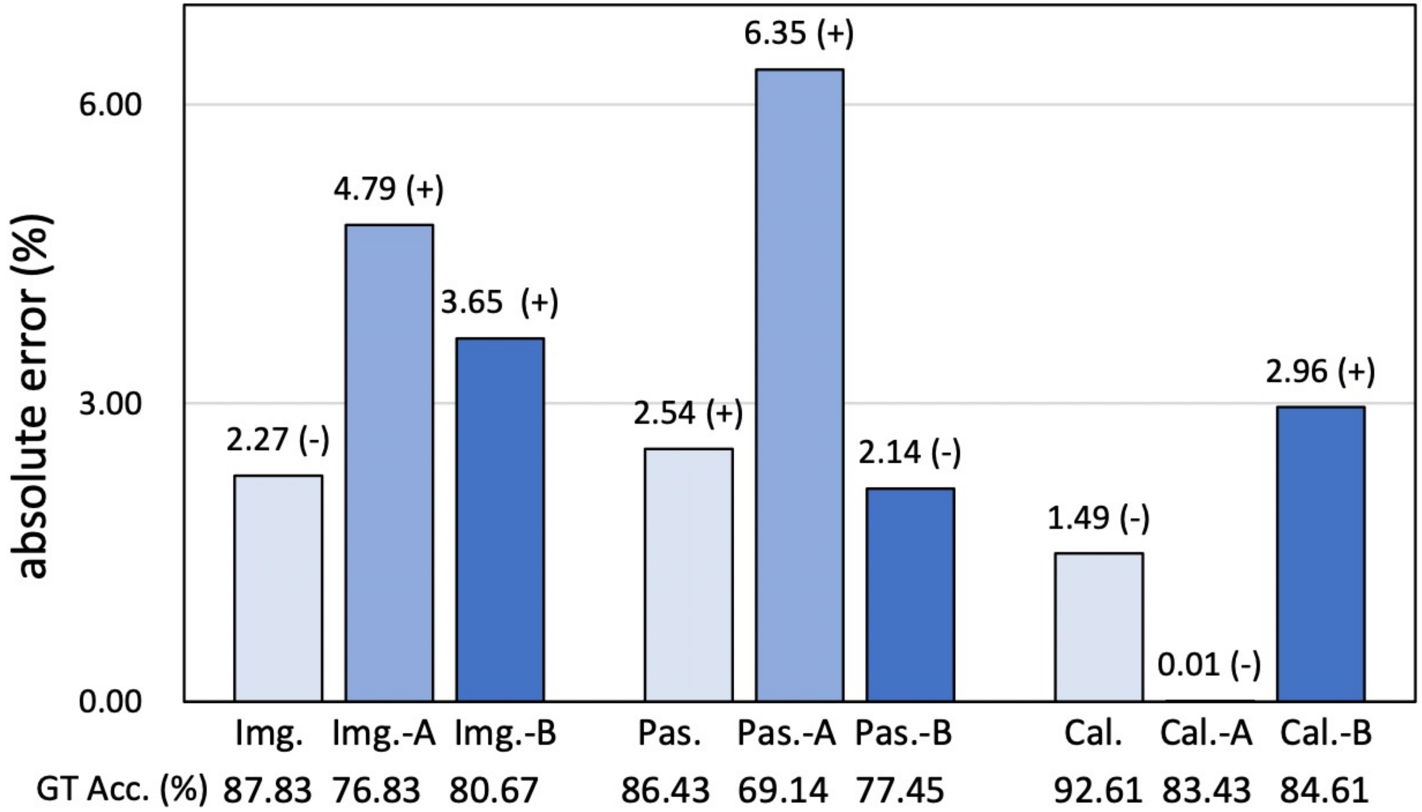
Train Set	MNIST			CIFAR-10		COCO			
Unseen Test Set	SVHN	USPS	RMSE↓	CIFAR-10.1	RMSE↓	Caltech	Pascal	ImageNet	RMSE↓
Ground-truth Accuracy	23.06	65.52	-	88.15	-	92.61	86.43	87.83	-
Prediction ($\tau_1 = 0.8$)	33.64	44.34	16.74	91.15	3.00	89.36	83.98	85.17	2.81
Prediction ($\tau_1 = 0.9$)	22.07	30.39	24.85	86.85	1.30	84.30	78.00	79.83	8.25
Entropy ($\tau_2 = 0.2$)	26.63	33.23	22.97	89.20	1.05	86.80	80.14	82.50	5.82
Entropy ($\tau_2 = 0.3$)	40.35	46.87	17.98	93.80	5.65	92.49	86.21	88.50	0.41
Linear Regression	24.84	53.10	8.87	91.89	3.74	90.70	89.29	90.98	2.68

Linear regression achieves reasonably good estimations on all test sets

Test sets undergo new transformations

- We add new image transformations to the test sets of COCO setup

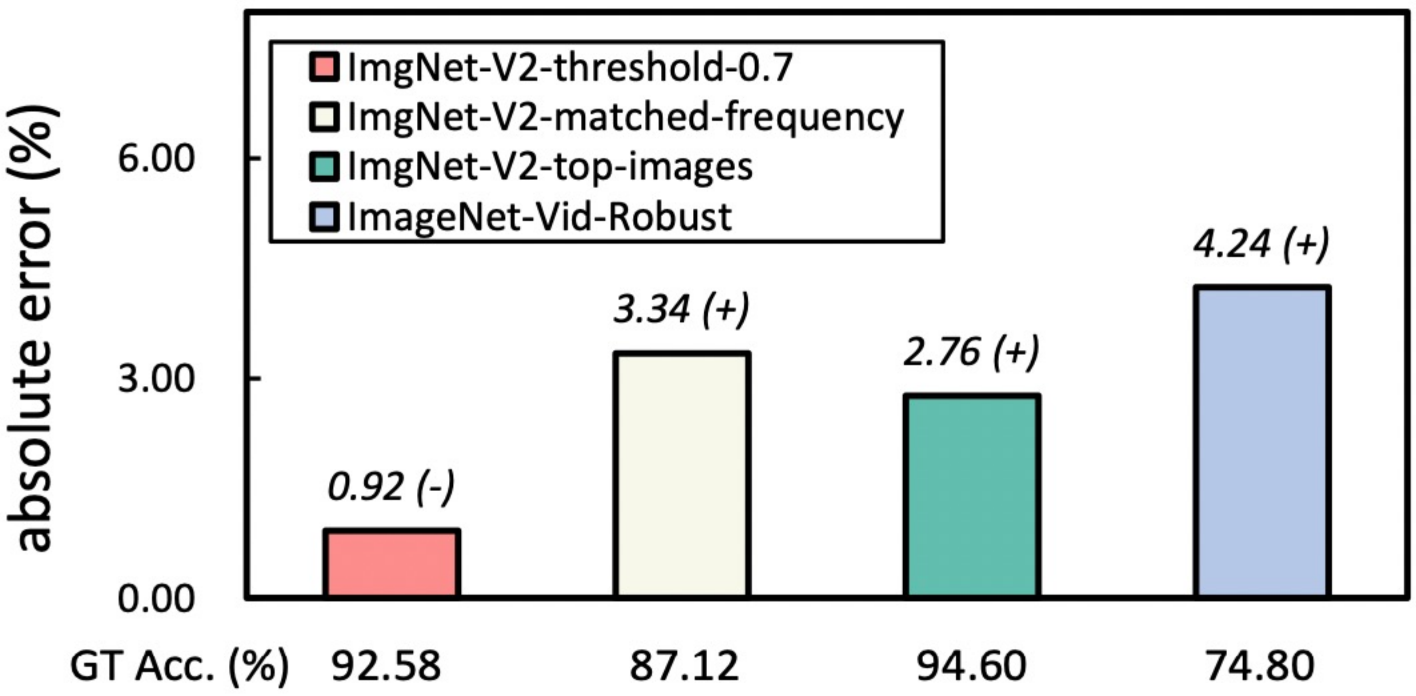
Random erasing, Posterize → *Group A*   *Pepper* and *FilterSmooth* → *Group B*



robust

More test sets under COCO setup

- We include more test sets to validate the generalization of regression model



generalizable

Conclusions and insights

- We study a very interesting problem:

  Evaluating model performance *without* ground truths

- We use a very simple method:

  Using accuracy of rotation prediction to

  estimate semantic classification accuracy

Conclusions and insights

- Limitation
  - Some corner cases (*e.g.*, balls and airplanes)
  - Rotation prediction should be well-defined and non- trivial
- Future Work
  - Use our correlation finding to select models without labels
  - Other machine learning tasks (*e.g.*, object detection)

Thank you!

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