SELF-SUPERVISED BENCHMARK LOTTERY ON IMAGENET:

DO MARGINAL IMPROVEMENTS TRANSLATE TO IMPROVEMENTS ON SIMILAR DATASETS?

Utku Ozbulak, Esla Timothy Anzaku, Solha Kang, Wesley De Neve, **Joris Vankerschaver (speaker)**

(Ghent University Global Campus, Incheon, South Korea)

GHENT UNIVERSITY
GLOBAL CAMPUS

Center for
Biosystems & Biotech Data Science

BENCHMARK LOTTERY

- Benchmark lottery (Deghani et al, 2021):
  - Performance of method changes just by changing task, dataset, …
  - ML evaluation is **fragile**
- Our contribution:
  - Assess performance of different **self-supervised learning** frameworks under different conditions

WHY SELF-SUPERVISED LEARNING (SSL)?

- Overcome shortage of labelled data
- Drawbacks:
  - Training is very compute intensive
  - Many different frameworks (> 100)

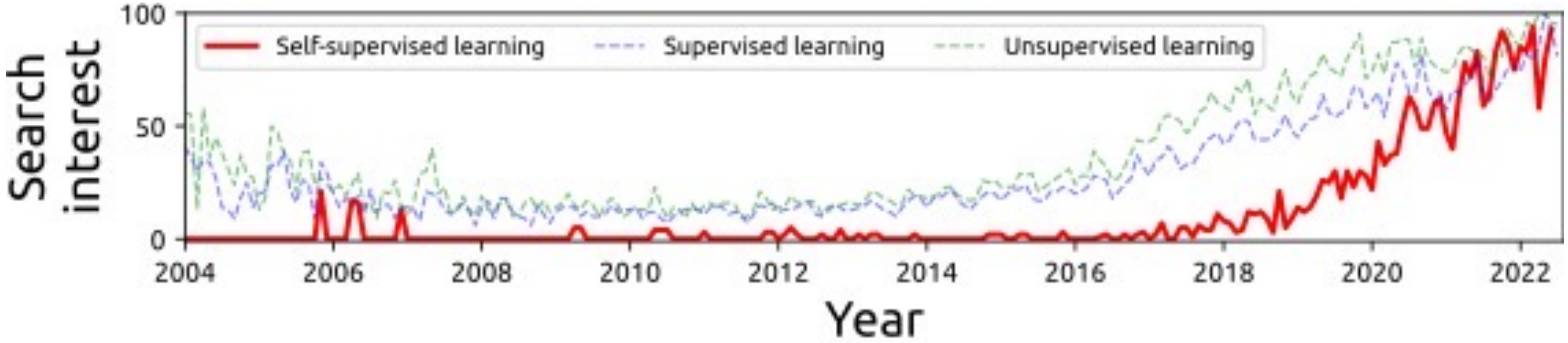
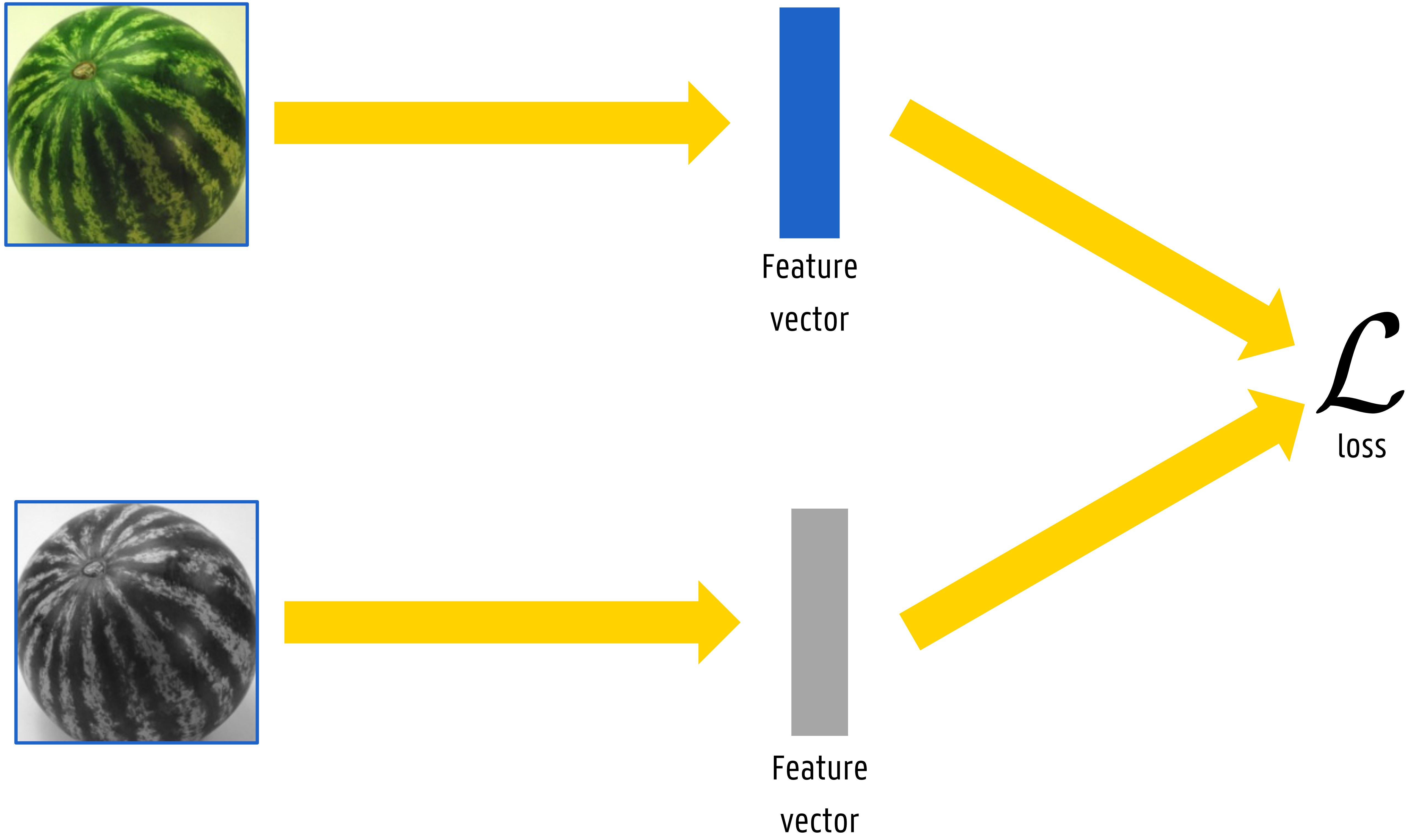


Image from Ozbulak et al: *Know your self-supervised learning.* Theory of Machine Learning Research (2023).

HOW DOES SELF-SUPERVISED LEARNING WORK?

Pretext task:

learn invariant representation of inputs



EXAMPLE PRETEXT TASKS

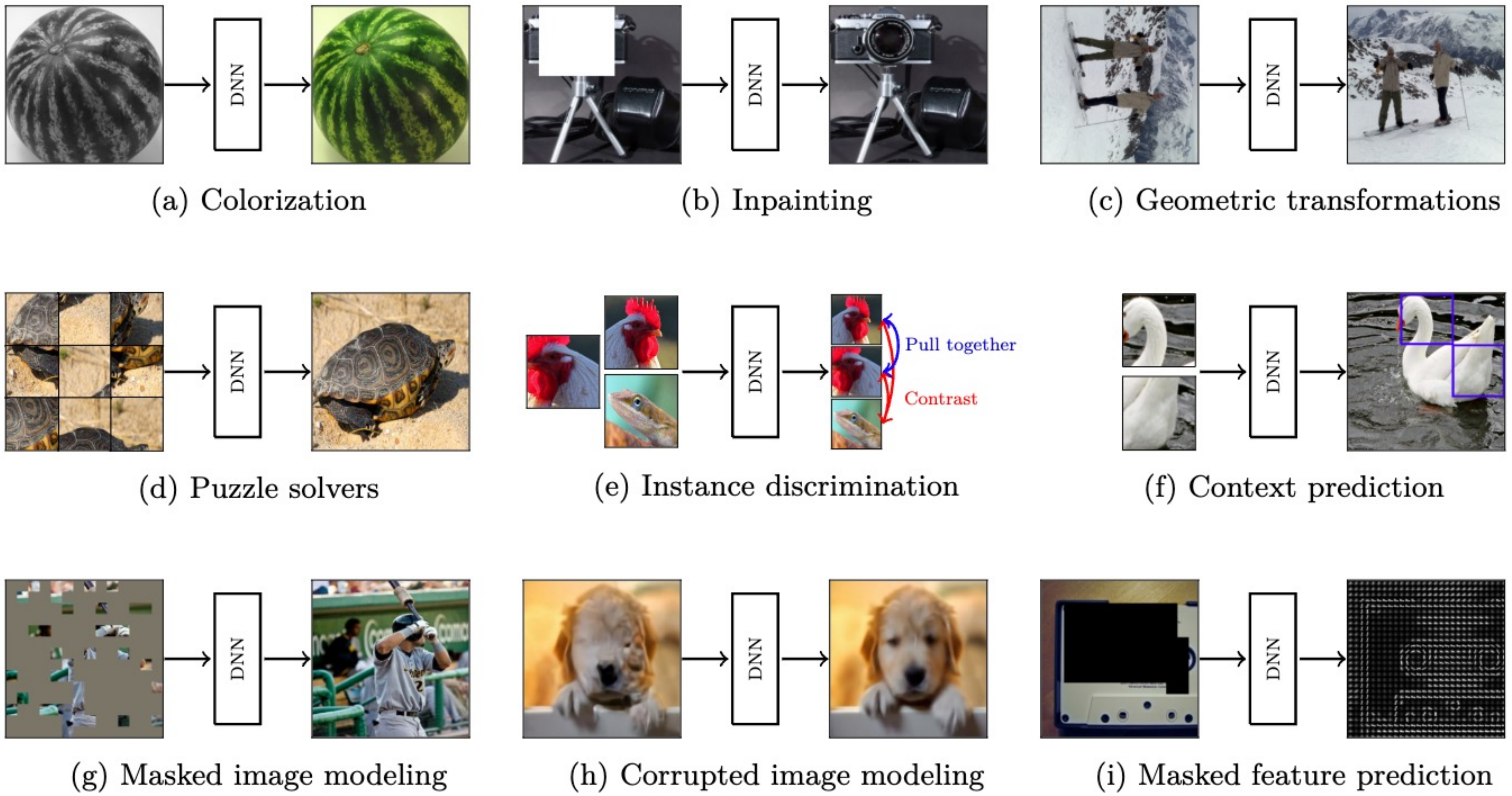
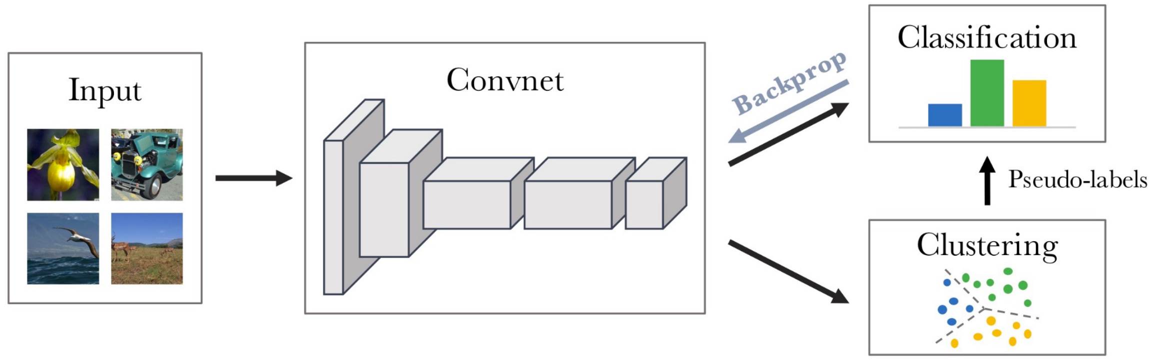


Image from Ozbulak et al: *Know your self-supervised learning.* Theory of Machine Learning Research (2023).

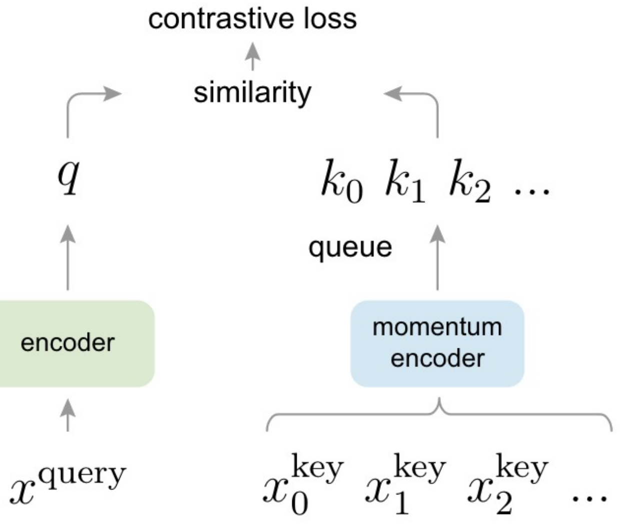
SSL FRAMEWORKS IN THIS STUDY

Clustering-based



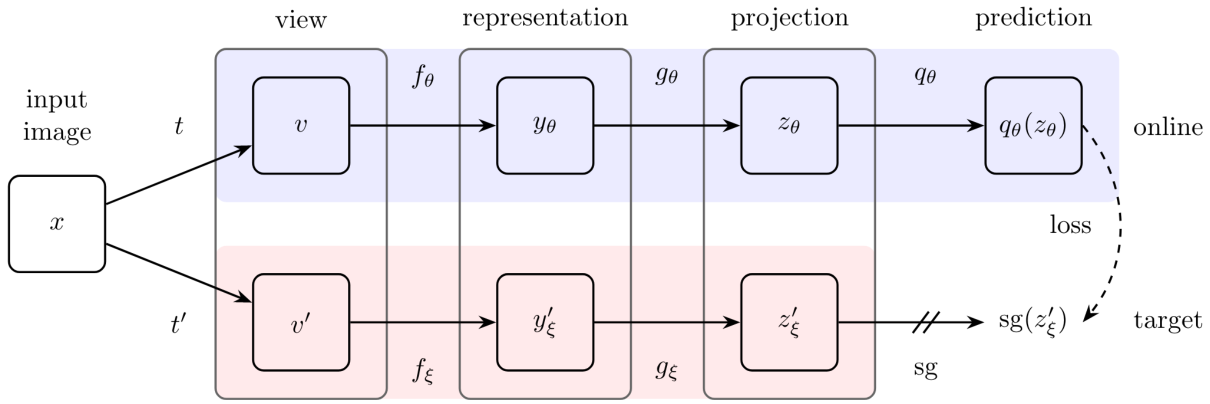
DeepC (2018), SeLa (2019), SwAV (2020)

Contrastive



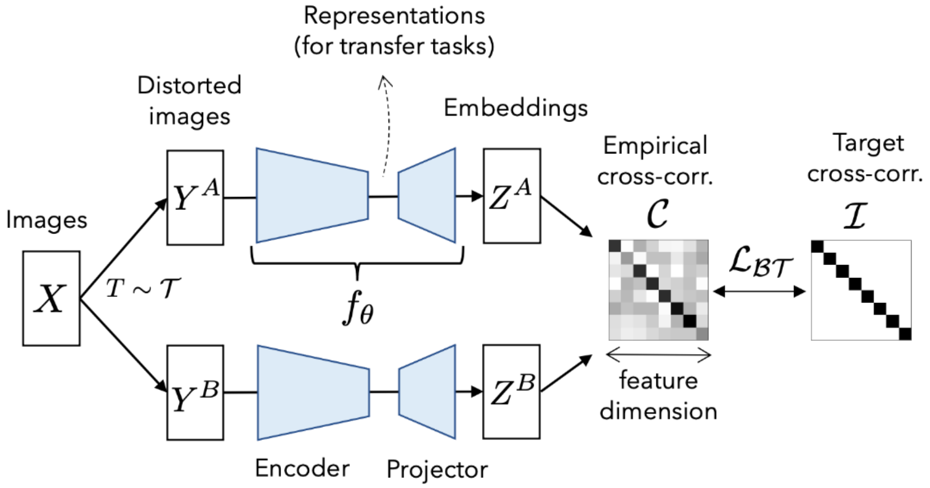
SimCLR (2020), MoCo (2020), PCL (2021)

Distillation



BYOL (2020), SimSiam (2020), DINO (2021), OBoW (2021)

Information Maximization



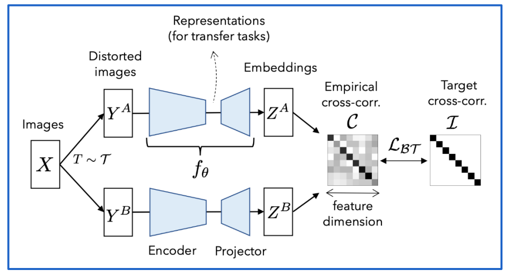
Barlow (2021), VicReg (2021)

EVALUATING AN SSL FRAMEWORK

On labelled dataset (e.g. ImageNet validation):

Pretext task

(features)

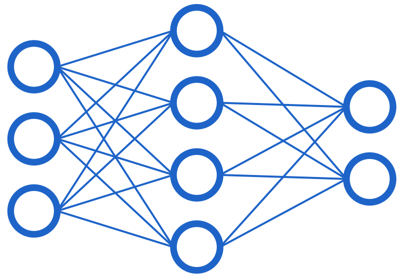


Feature

vector

Downstream task

(classification)



melon

Evaluation metrics

- kNN evaluation
- Linear evaluation
- Fine-tuning

GHENT UNIVERSITY
GLOBAL CAMPUS

DATASETS: IMAGENET VARIANTS

ImageNet Rendition

Out-of-distribution generalization

ImageNet Sketch

Global vs. local features

ImageNet ReaL

Multi-label classification

ImageNet Adversarial

Adversarial examples

(misclassified examples)

Dataset	Image count	Classes	Image per class	Multi-label
Validation	50,000	1,000	50	✗
ReaL	50,000	1,000	50	✓
v2	10,000	1,000	10	✗
Rendition	30,000	200	∼150	✗
Sketch	50,889	1,000	∼50	✗
Adversarial	7,500	200	∼37	✗

ImageNet v2

METHODOLOGY

- 12 SSL frameworks
- ImageNet + 5 variants
- Approach:
  - Train SSL framework on ImageNet
  - Verify accuracy (within 1% of paper)
  - Perform linear evaluation on datasets

EXPERIMENTAL RESULTS: ACCURACY

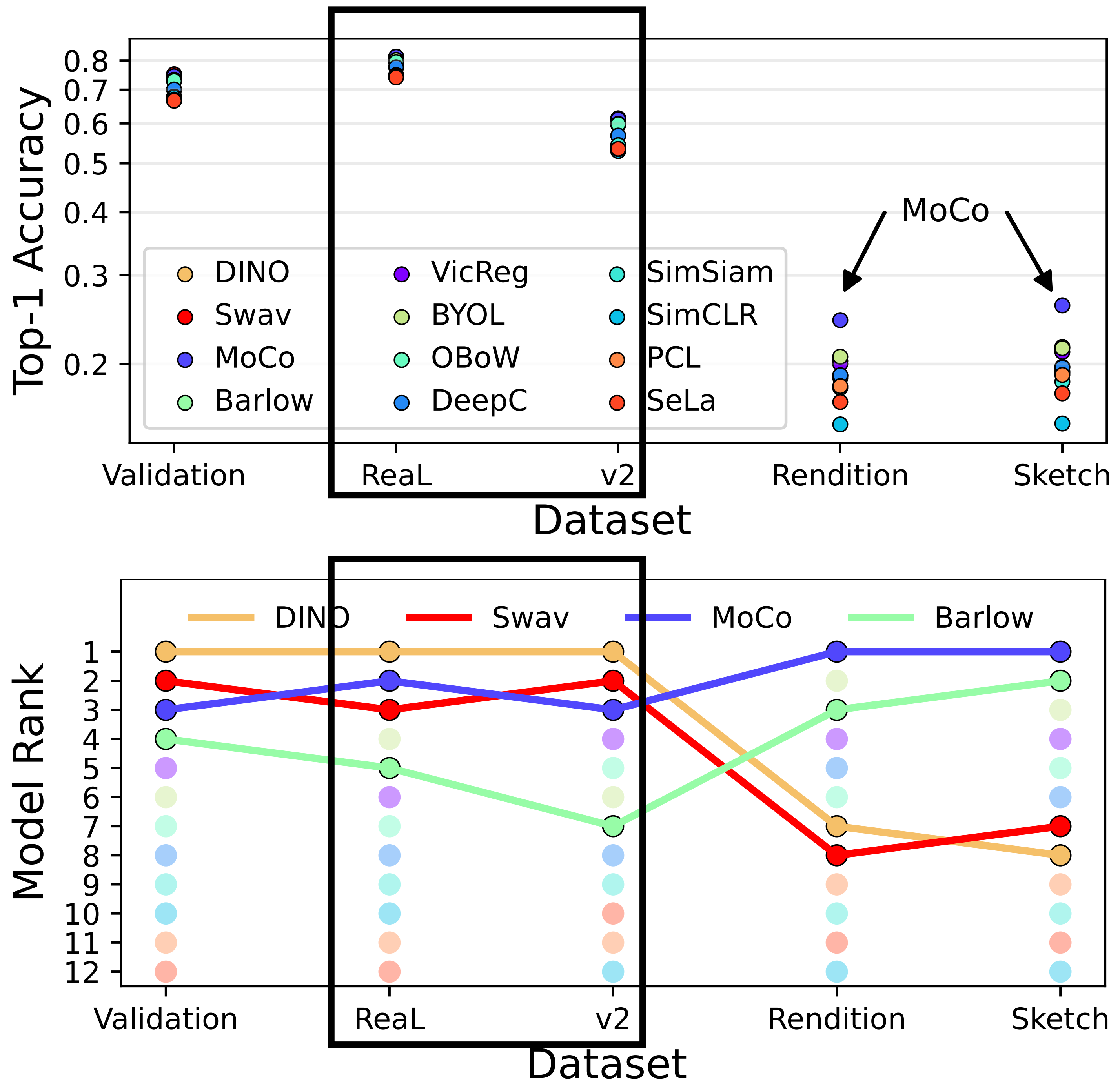
Accuracy on **Real**:

- Similar accuracy
- Small ranking differences

👉 Robustness of models

Accuracy on **v2**:

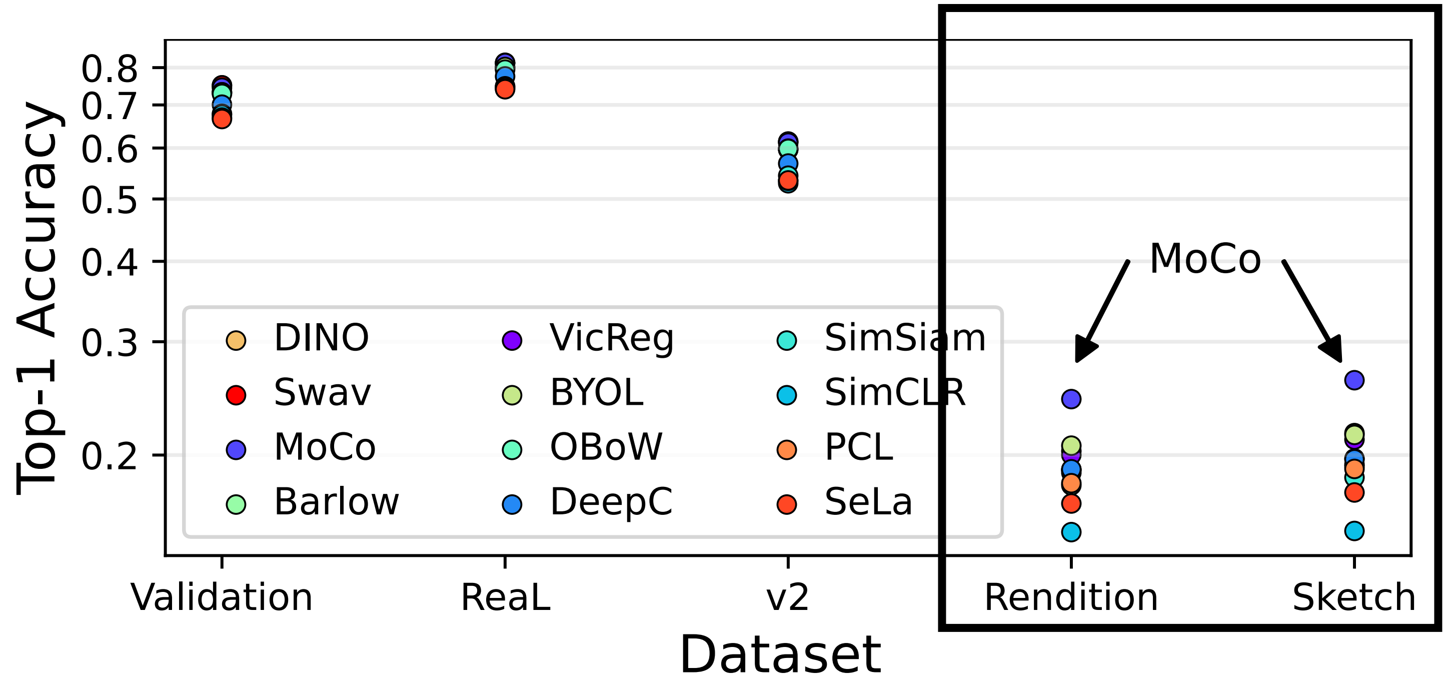
- Drop in accuracy of 10-15%
- Small ranking differences

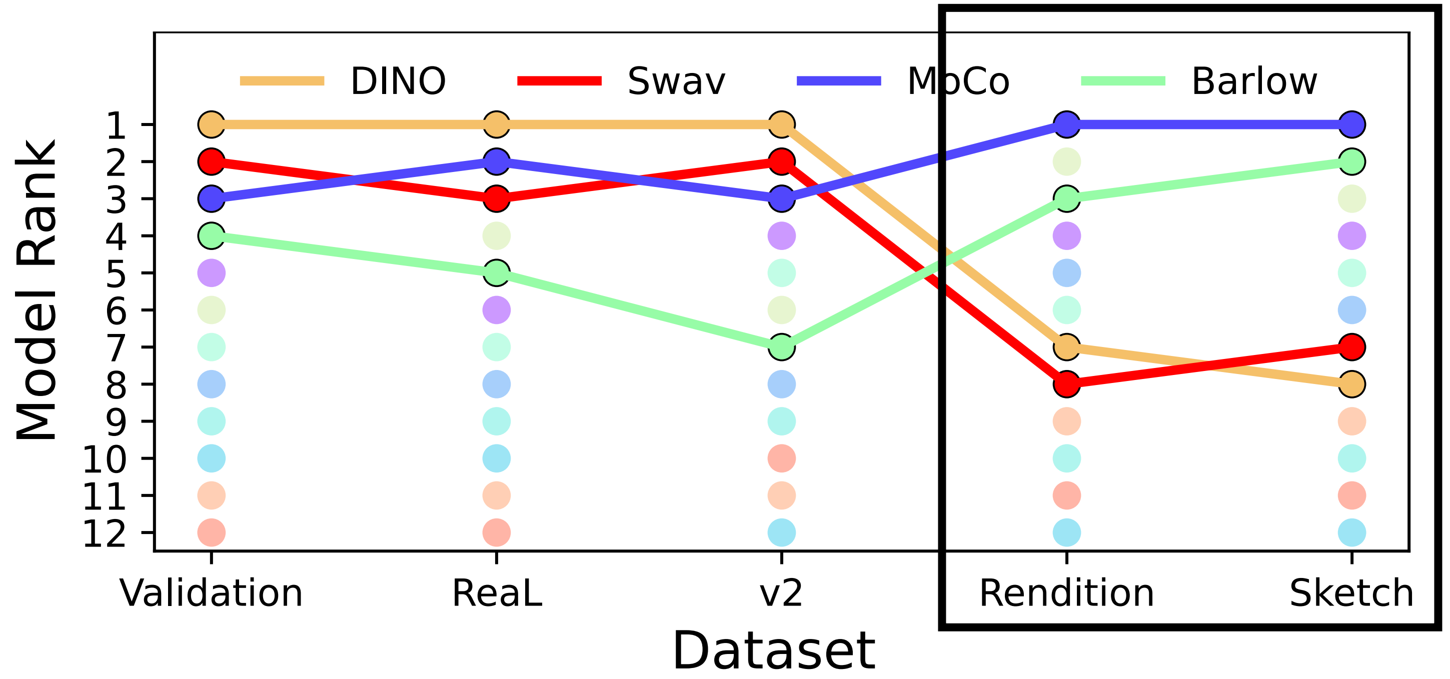


EXPERIMENTAL RESULTS: ACCURACY

Accuracy on **Rendition/Sketch**:

- Significant decline
- Large changes in ranking
- MoCo and Barlow are robust
- MoCo is strong all-round





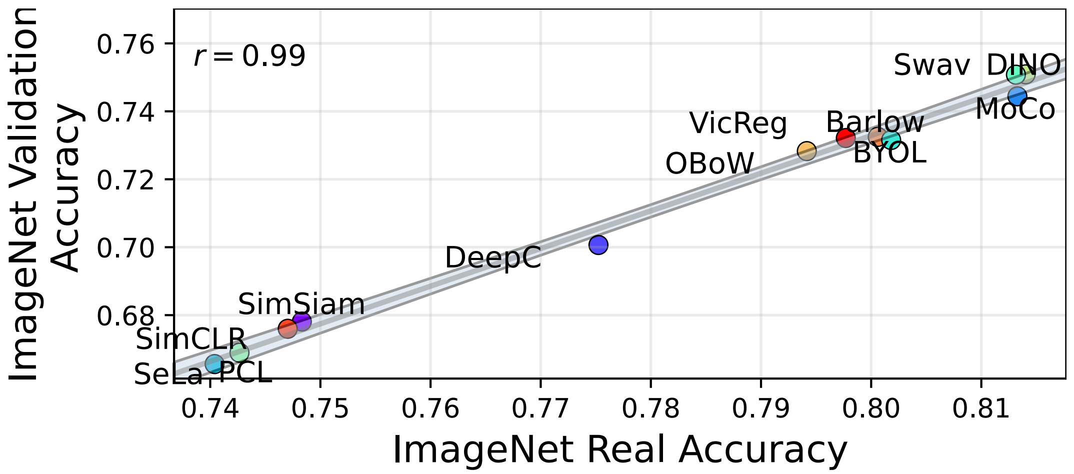
GHENT UNIVERSITY
GLOBAL CAMPUS

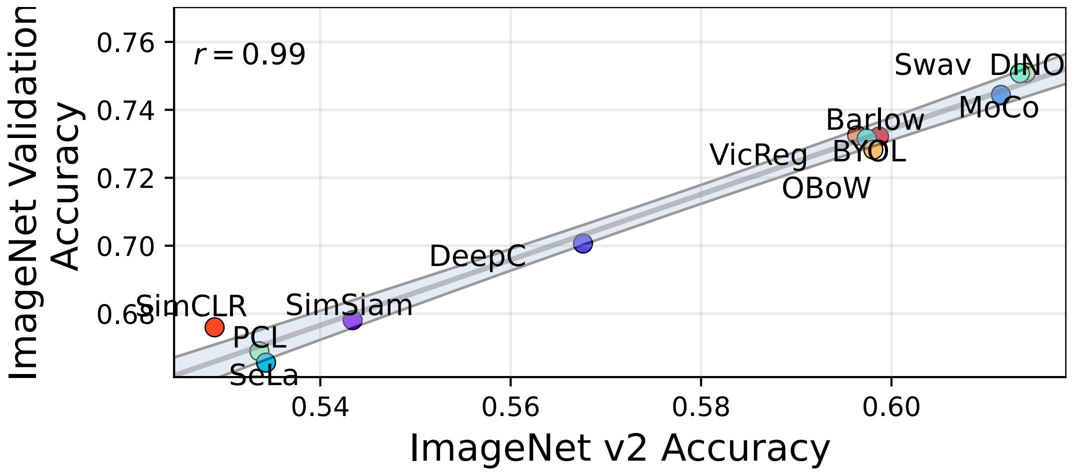
EXPERIMENTAL RESULTS: ACCURACY

Accuracy on **Adversarial**: single digits 💀

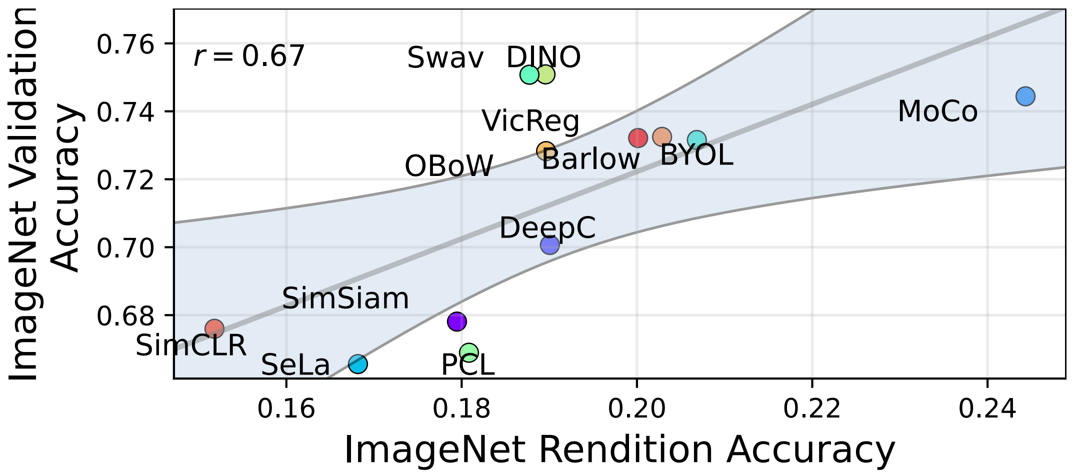
Model	Val.	ReaL	v2	Rendition	Sketch	Adv.
DINO	75.1	81.4	61.4	18.9	19.2	2.3
Swav	75.0	81.3	61.3	18.7	19.3	2.4
MoCo	74.4	81.3	61.1	24.4	26.1	1.8
Barlow	73.2	80.0	59.6	20.2	21.6	1.5
VicReg	73.2	79.7	59.8	20.0	21.1	1.6
BYOL	73.1	80.1	59.7	20.6	21.4	1.6
OBoW	72.8	79.4	59.8	18.9	19.7	3.3
DeepC	70.0	77.5	56.7	19.0	19.6	1.4
SimSiam	67.8	74.8	54.3	17.9	18.4	1.2
SimCLR	67.5	74.7	52.8	15.1	15.2	1.1
PCL	66.8	74.2	53.3	18.0	19.0	1.3
SeLa	66.5	74.0	53.4	16.8	17.4	1.2

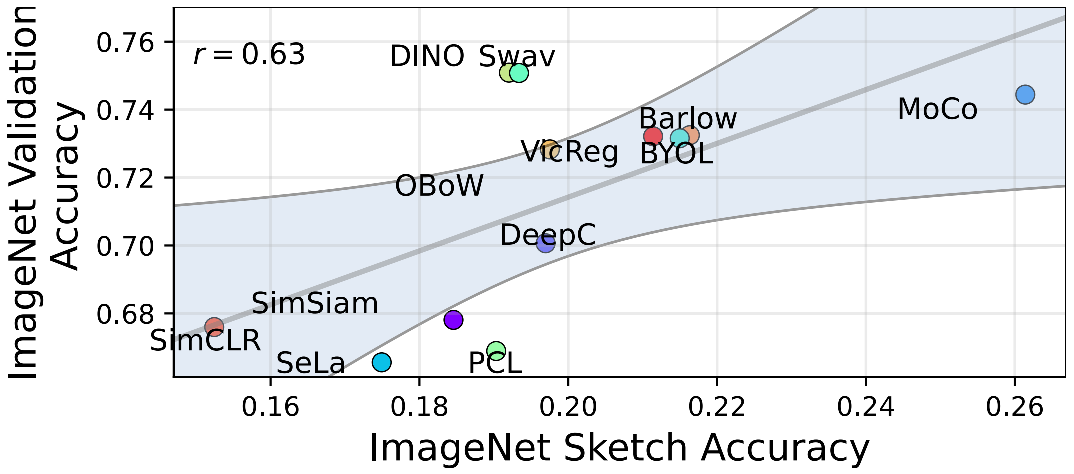
EXPERIMENTAL RESULTS: CORRELATION WITH IMAGENET





Strong correlation with Real/v2: similar datasets, strong generalizability





Weaker correlation with Rendition/Sketch: lack of OOD generalization

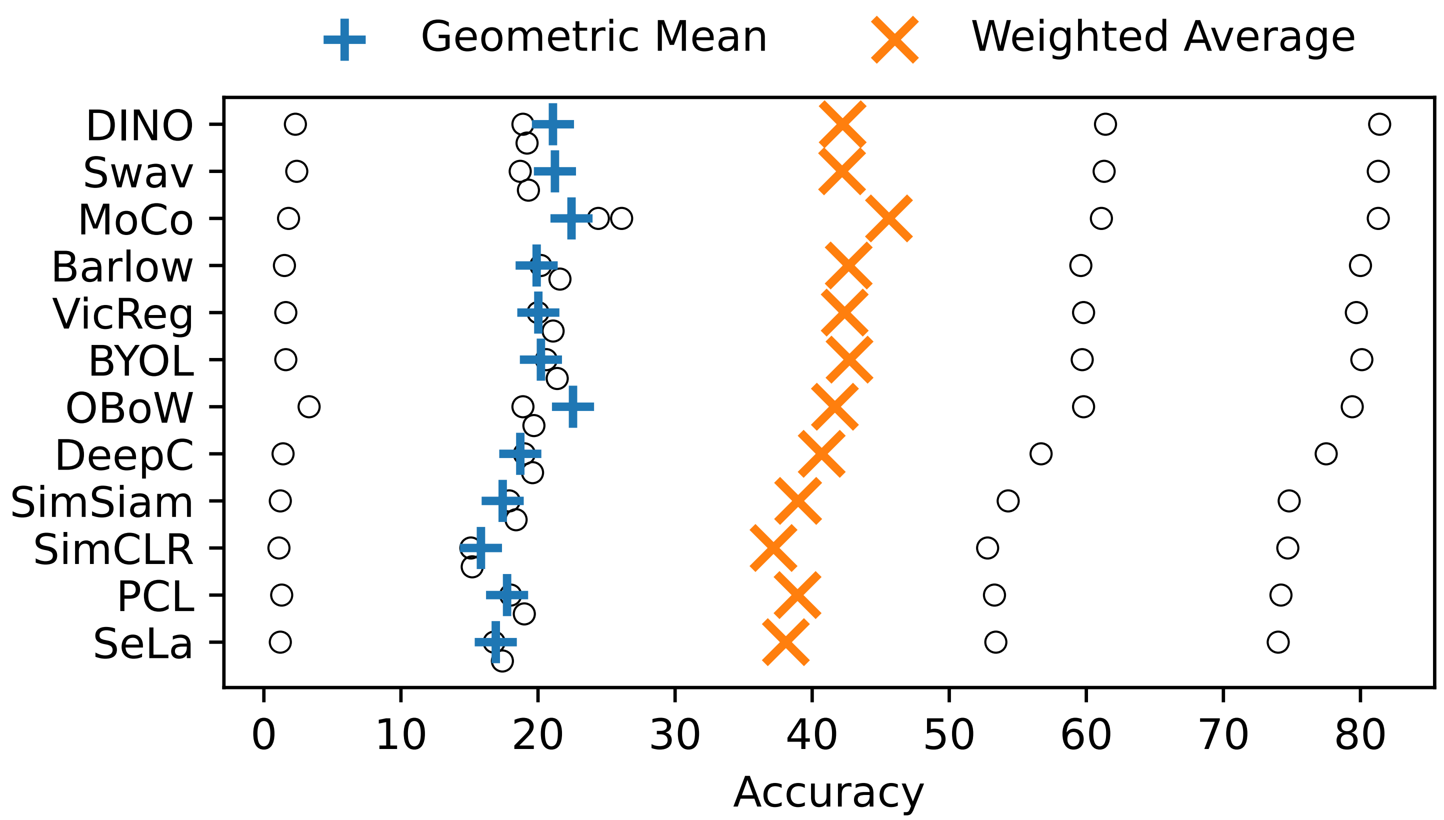
EXPERIMENTAL RESULTS: AGGREGATE PERFORMANCE

Weighted average: prioritizes **large datasets**

– MoCo does best

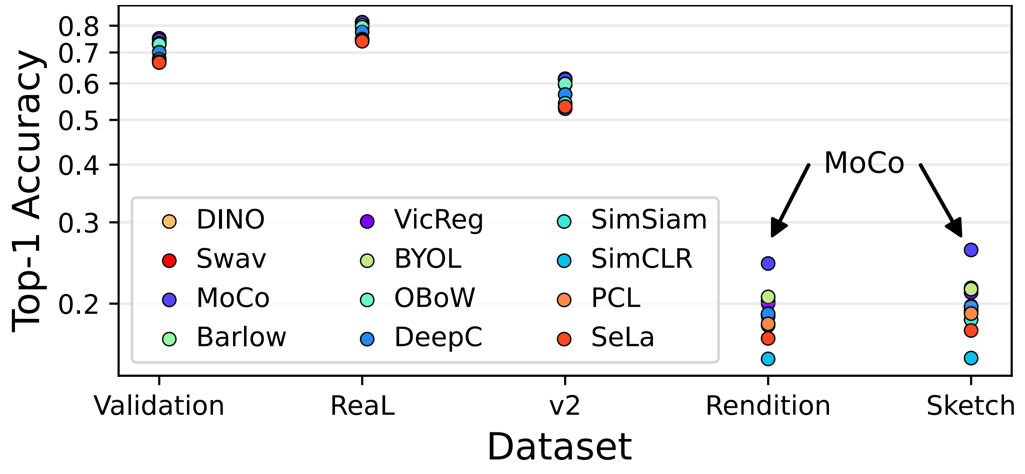
Geometric mean: prioritizes **worst-performing** datasets (pessimistic)

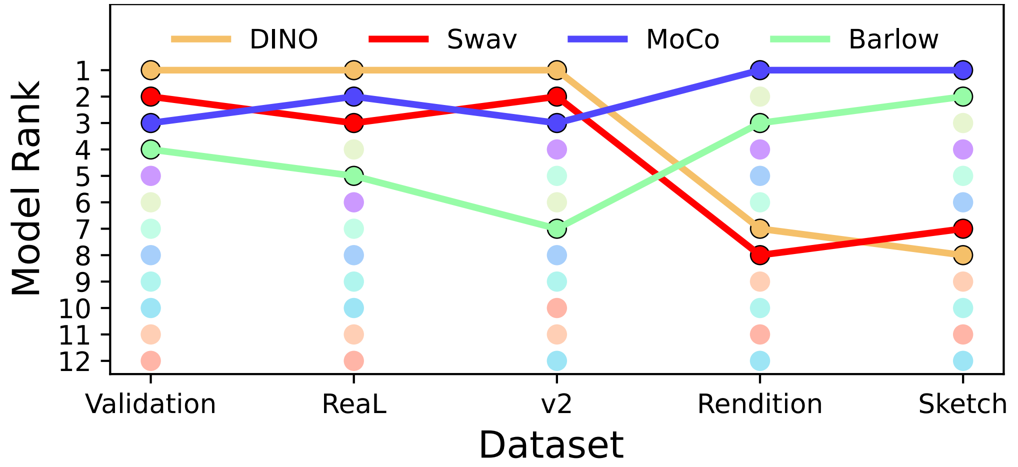
– MoCo/OBoW do well (adv)



CONCLUSIONS

- Evalution on ImageNet only is misleading
- Different datasets bring out different aspects
  - Out-of-distribution generalization
- MoCo is a strong all-round contender





Thank you for your time!

Contact:

- Utku.Ozbulak@ghent.ac.kr
- Joris.Vankerschaver@ghent.ac.kr

GHENT UNIVERSITY
GLOBAL CAMPUS