Masked Image Modeling as Vision Pre-training

Methodology, Understanding and Data-scaling Capability

Yue Cao

Microsoft Research Asia

June 2nd, 2022

@ BAAI

LeCun's Cake Analogy

- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

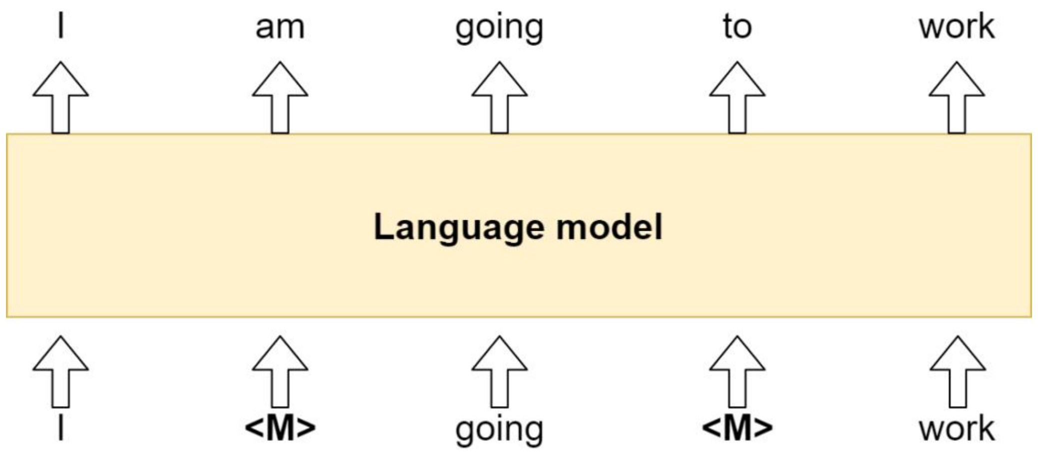
Why is it so important?

- Baby learns how the world works primarily by observation
- Unlimited data without annotations could be utilized


Photos courtesy of Emmanuel Dupoux

Credit by Yann LeCun

Pre-training paradigm of NLP

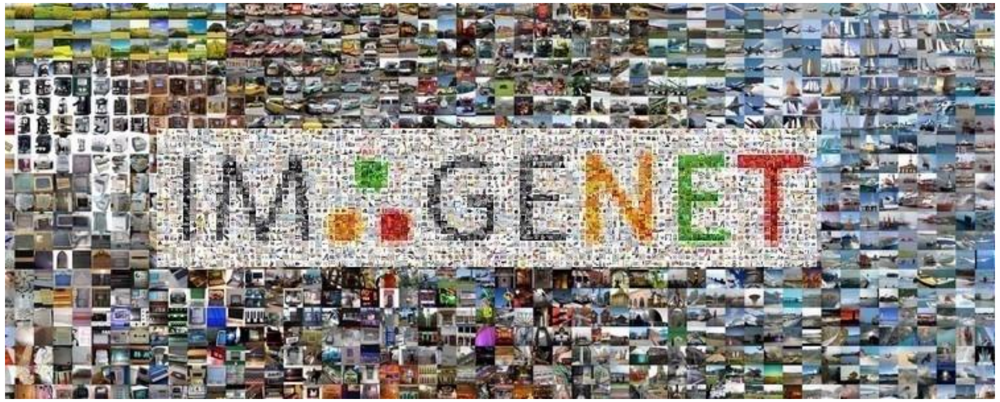


Self-supervised Pre-training with Masked Language Modeling

Fine-tuning

- Question Answering (SQuAD)
- Commonsense Reasoning (SWAG)
- Text Summarization
- Sentiment Classification
- ……

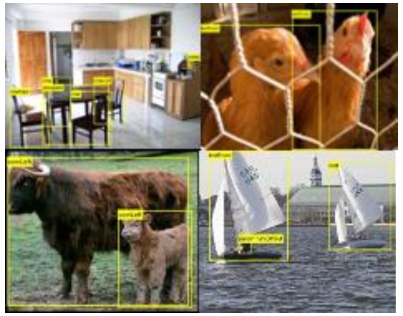
Pre-training paradigm of CV



Supervised classification on ImageNet-1K
as pre-training

Fine-tuning

Semantic Segmentation

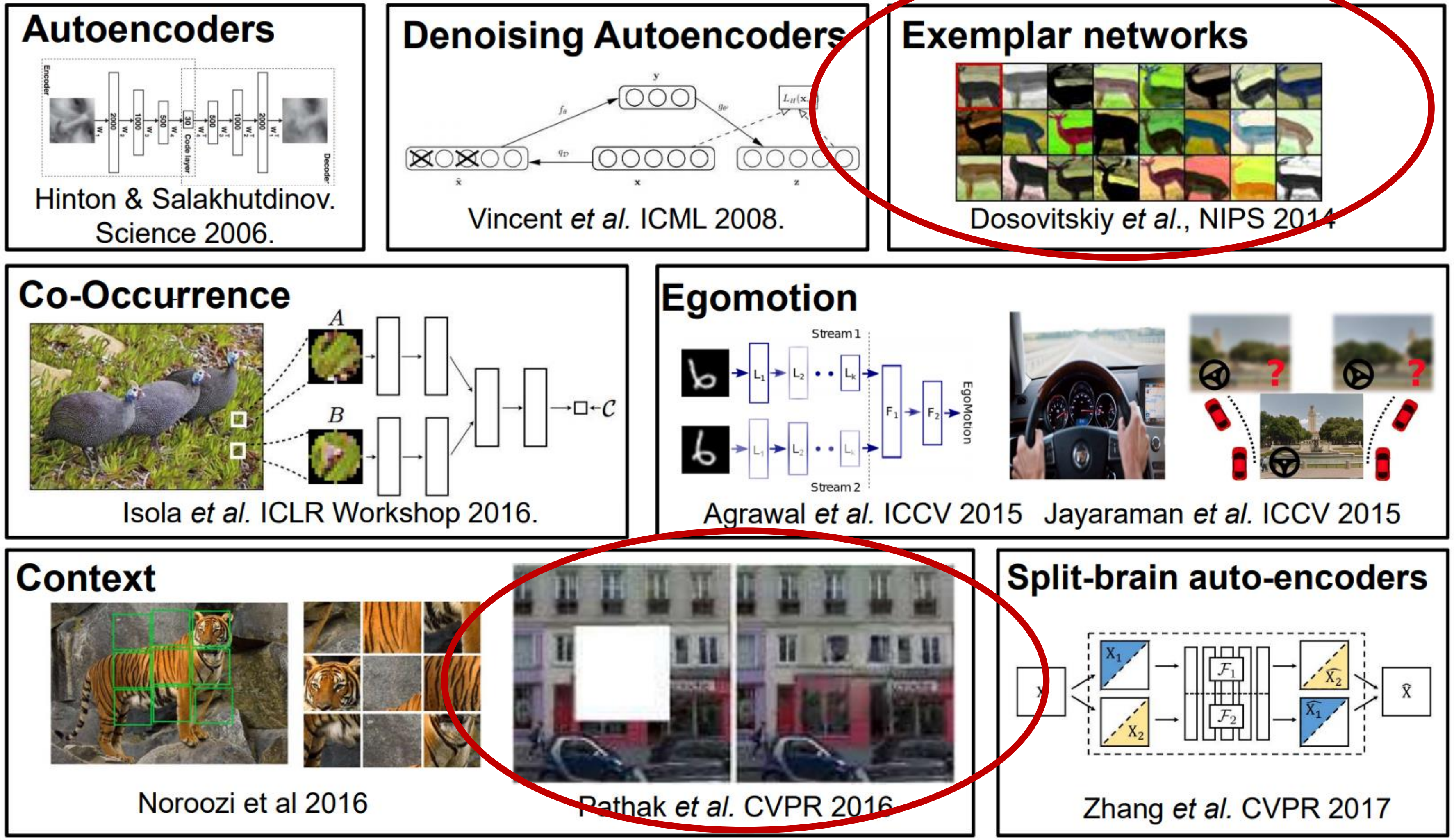


Object Detection

Fine-grained Image
Classification

Explorations of Self-supervised Learning in Vision

Credit by Andrew Zisserman



Milestone

- Self-supervised Pre-training + Fine-tuning

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Code: https://github.com/facebookresearch/moco

2019.11

MoCo

Facebook AI Research

- 在7个下游任务上，**自监督**预训练**首次超越**有监督预训练

Discriminative Pre-training

Task: Distinguish each image

2014.6
Exemplar CNNs
Univ. of Freiburg

2018.5
Memory bank
UC Berkeley

2018.12
Deep metric transfer
MSRA

2019.11
MoCo
FAIR

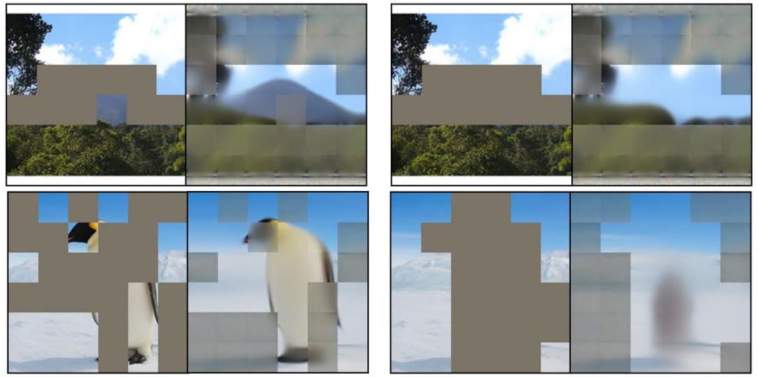
2020.2
SimCLR
Google Brain

2020.6
BYOL, SwAV
DeepMind, FAIR

2020.6
PIC
MSRA

2020.11
SimSiam, PixPro
FAIR, MSRA

Generative Pre-training



Task: Predict the masked area

2016.4
Context Encoder
UC Berkeley

2018.7
Contrastive Predictive Coding
DeepMind/UC Berkeley

2020.6
iGPT
OpenAI

2020.10
ViT
Google Brain

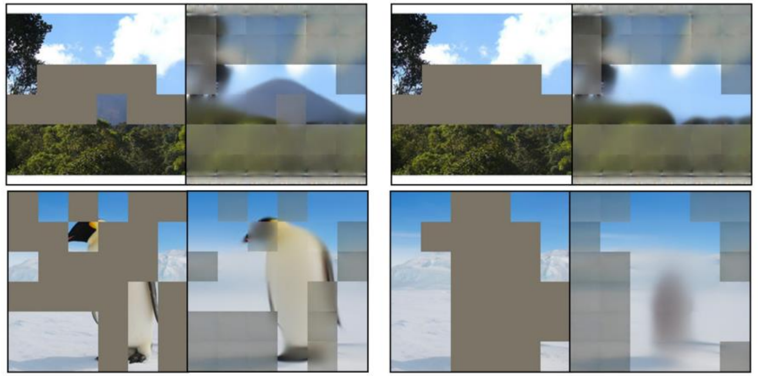
2021.6
BEiT
MSRA

2021.10
MAE
FAIR

2021.11
SimMIM
MSRA

Masked Image Modeling

- Could MIM be simple but effective?
- How and where does MIM pretraining work?
- Could MIM benefit from larger-scale data?

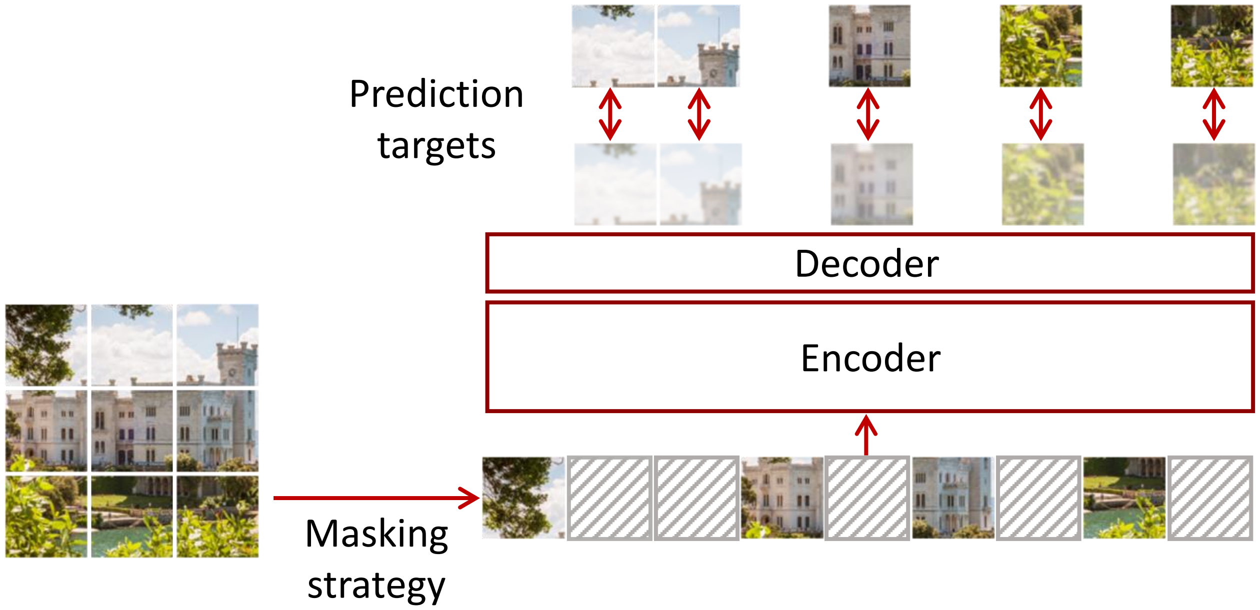


Task: Predict the masked area

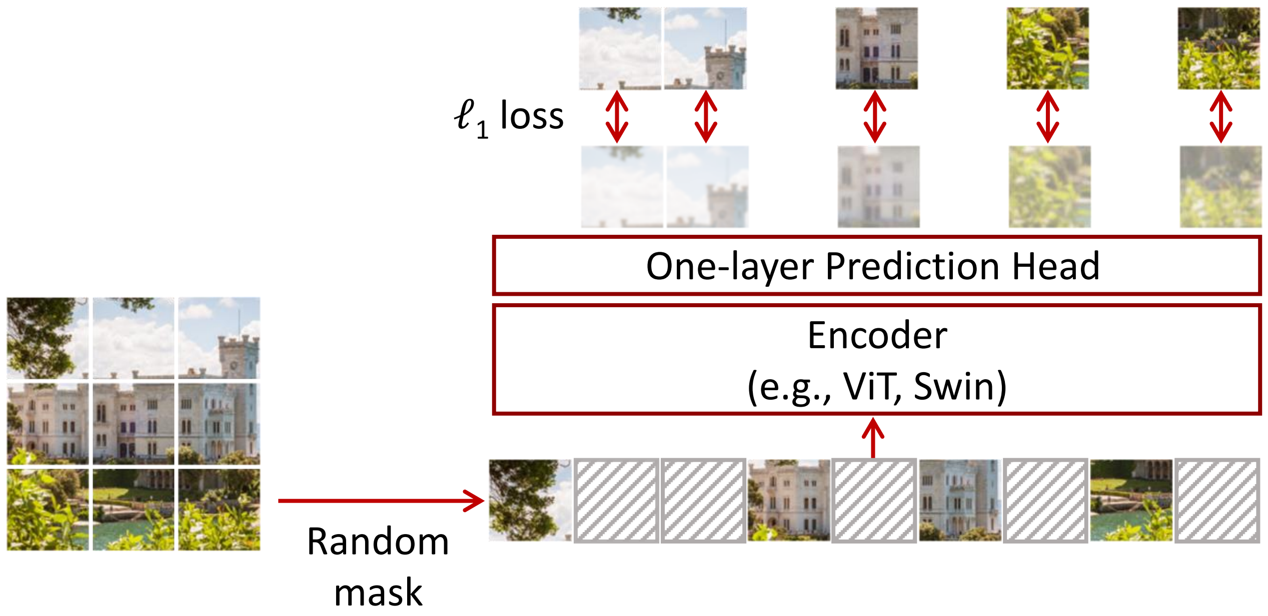
Masked Image Modeling

- **Could MIM be simple but effective?**
  - **SimMIM: A Simple Framework for Masked Image Modeling, CVPR 2022**
- How and where does MIM work?
  - Revealing the Dark Secrets of Masked Image Modeling, Arxiv 2022
- Could MIM benefit from larger-scale data?
  - On Data Scaling in Masked Image Modeling, Arxiv 2022

SimMIM: A Simple Framework on MIM



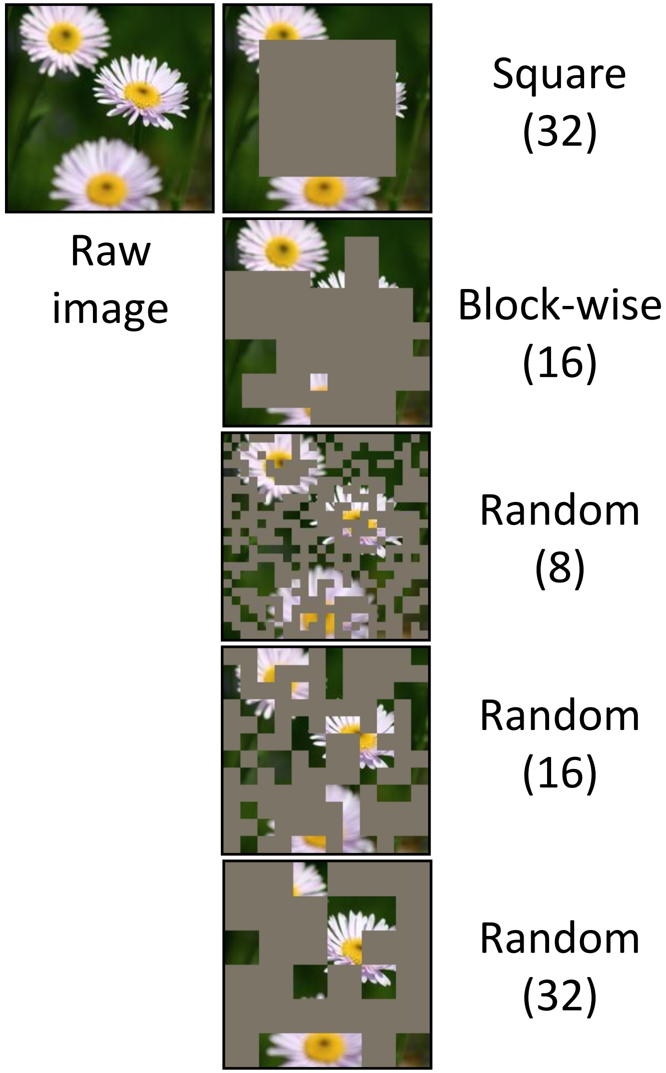
SimMIM: A Simple Framework on MIM



(a) **Masking strategy**: Random masking with relatively large patch size (e.g., 32x32)
(b) **Prediction heads**: An extremely lightweight prediction head (e.g., a linear layer)
(c) **Prediction targets**: A simple raw pixel regression task
(d) **Encoder architectures**: ViT, Swin and ResNet could all benefit from SimMIM

- (a) Masking strategy
- (b) Prediction heads
- (c) Prediction targets
- (d) Encoder types

Ablation: Masking Strategy



- **A simple random masking** works well
- **Large patch size/High mask rate** matters
  - Visual signals are redundant spatially and exhibit strong locality

	16/32	0.8	82.4/82.5
	4/8/16/32	0.4	81.9/82.0/82.4/82.9
	4/8/16/32	0.6	82.0/82.1/82.7/82.8
random	4/8/16/32	0.8	82.1/82.4/82.8/82.4
	64	0.1	82.6
	64	0.2	82.6
	32	0.1	82.7
	32	0.2	82.8
	32	0.3	82.8
	32	0.4	82.9
random	32	0.5	**83.0**
	32	0.6	82.8
	32	0.7	82.7
	32	0.8	82.4
	32	0.9	82.4

Ablation: Prediction Heads

Head	#params	Training costs	Top-1 acc (%)
Linear	89.9M	1×	82.8
2-layer MLP	90.9M	1.2×	82.8
inverse Swin-T	115.2M	1.7×	82.4
inverse Swin-B	174.8M	2.3×	82.5

- An extremely **lightweight prediction head** (e.g., a linear layer) achieves similarly or slightly better performance than that of heavier prediction heads

Ablation: Prediction Targets

Loss	Pred. Resolution	Top-1 acc (%)
Classification		
8-bin	192^2	82.7
8-bin	48^2	82.7
256-bin	192^2	N/A
256-bin	48^2	82.3
iGPT cluster	192^2	N/A
iGPT cluster	48^2	82.4
BEiT	-	82.7
Regression		
ℓ_2	192^2	82.7
smooth-ℓ_1	192^2	82.7
ℓ_1	192^2	82.8
ℓ_1	48^2	82.7
ℓ_1	6^2	82.3

Table 5. Ablation on different prediction targets.

- A **simple raw pixel regression task** performs no worse than the specialized classification approaches, such as tokenization (BEiT), clustering (iGPT), or discretization (ViT)
    - The visual signal is continuous in nature

(a) Masking strategy
(b) Prediction heads
(c) Prediction targets
(d) Encoder types

Ablation: Encoder

Methods	Input Size	Fine-tuning Top-1 acc (%)	Linear eval Top-1 acc (%)	Pre-training costs
Sup. baseline [46]	224^2	81.8	-	-
DINO [5]	224^2	82.8	78.2	2.0×
MoCo v3 [9]	224^2	83.2	76.7	1.8×
ViT [15]	384^2	79.9	-	~4.0×
BEiT [1]	224^2	83.2 +0.6	56.7	1.5×†
Ours	224^2	**83.8**	56.7	1.0×

Table 6. System-level comparison using ViT-B as the encoder. Training costs are counted in relative to our approach. † BEiT requires an additional stage to pre-train dVAE, which is not counted.

ResNet-50x4	Input Size	Fine-tuning top-1 acc
Sup. baseline	224	80.7 +0.9
Ours	224	**81.6**

- ViT, Swin and ResNet could all benefit from SimMIM

System-level Comparison

Methods	Pre-train	Fine-tune	Backbone	Top-1 acc (%)	Param
Sup.	192^2	224^2	Swin-B	83.3	88M
Sup.	192^2	224^2	Swin-L	83.5	197M
Sup.	192^2	224^2	SwinV2-H	83.3	658M
Ours	192^2	224^2	Swin-B	84.0 +0.7	88M
Ours	192^2	224^2	Swin-L	85.4 +1.9	197M
Ours	192^2	224^2	SwinV2-H	85.7 +2.4	658M
Ours	192^2	512^2	SwinV2-H	87.1	658M
Ours	192^2	640^2	SwinV2-G	90.2	3.0B

Table 7. Scaling experiments with Swin Transformer as backbone architectures. All our models are pre-trained with input of 192^2. Different to other models, Swin-G is trained on a privately collected ImageNet-22K-ext dataset, with details described in [33].

Backbone	Sup.		Ours	
	COCO mAP^{box}	ADE20K mIoU	COCO mAP^{box}	ADE20K mIoU
Swin-B	50.2	50.4	52.3 +2.1	52.8 +2.4
Swin-L	50.9	50.0	53.8 +2.9	53.5 +3.5
SwinV2-H	50.2	49.8	54.4 +4.2	54.2 +4.4

Table 12. Scaling experiments with Swin on COCO and ADE20K.

Averaged Distance

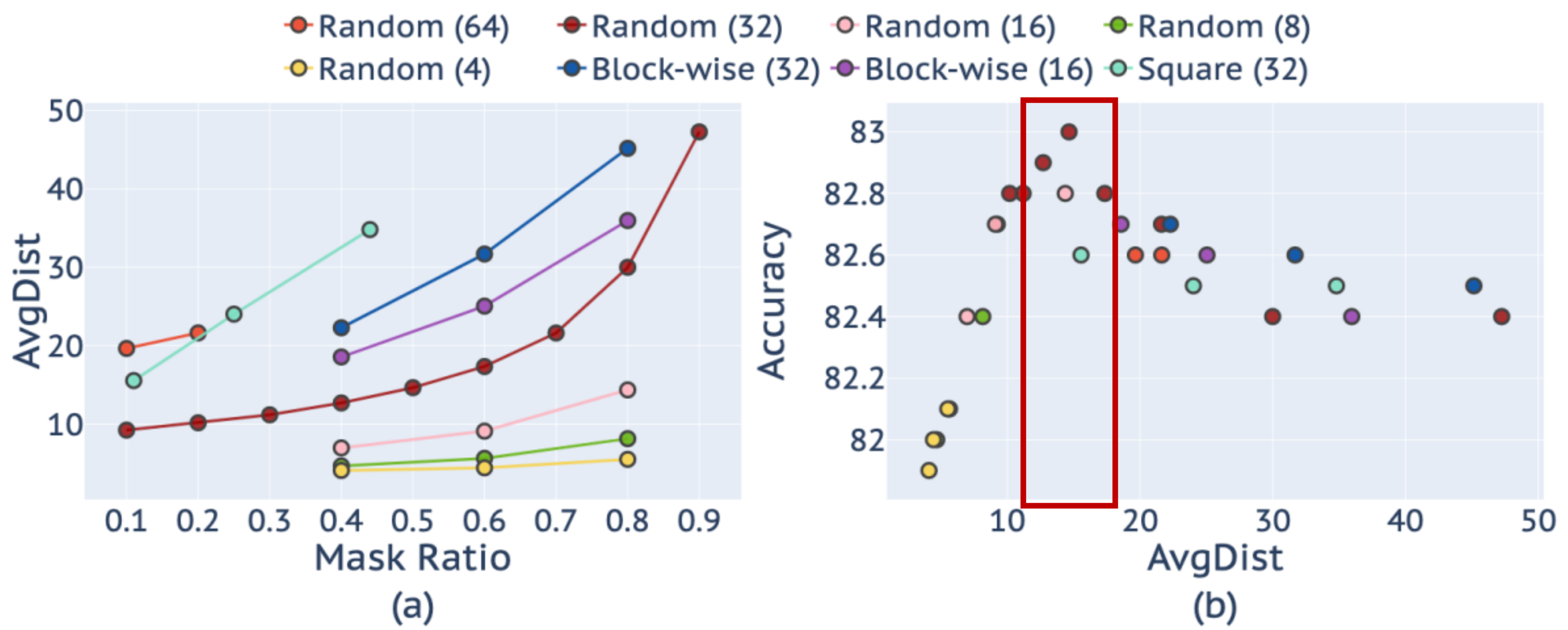
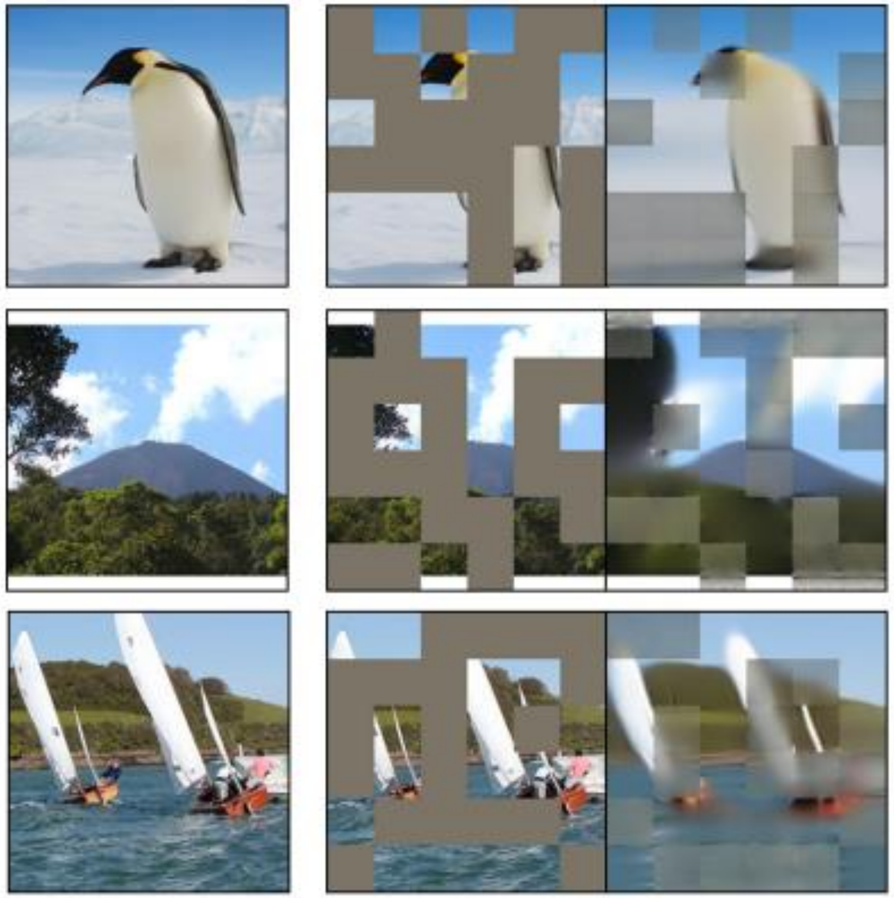


Figure 3. (a) *AvgDist* (averaged distance of masked pixels to the nearest visible pixels) w.r.t. different masking ratios using different masking strategies and different masked patch sizes; (b) fine-tuning performance (top-1 accuracy) w.r.t. *AvgDist*.

Visualizations



Masked Image Modeling

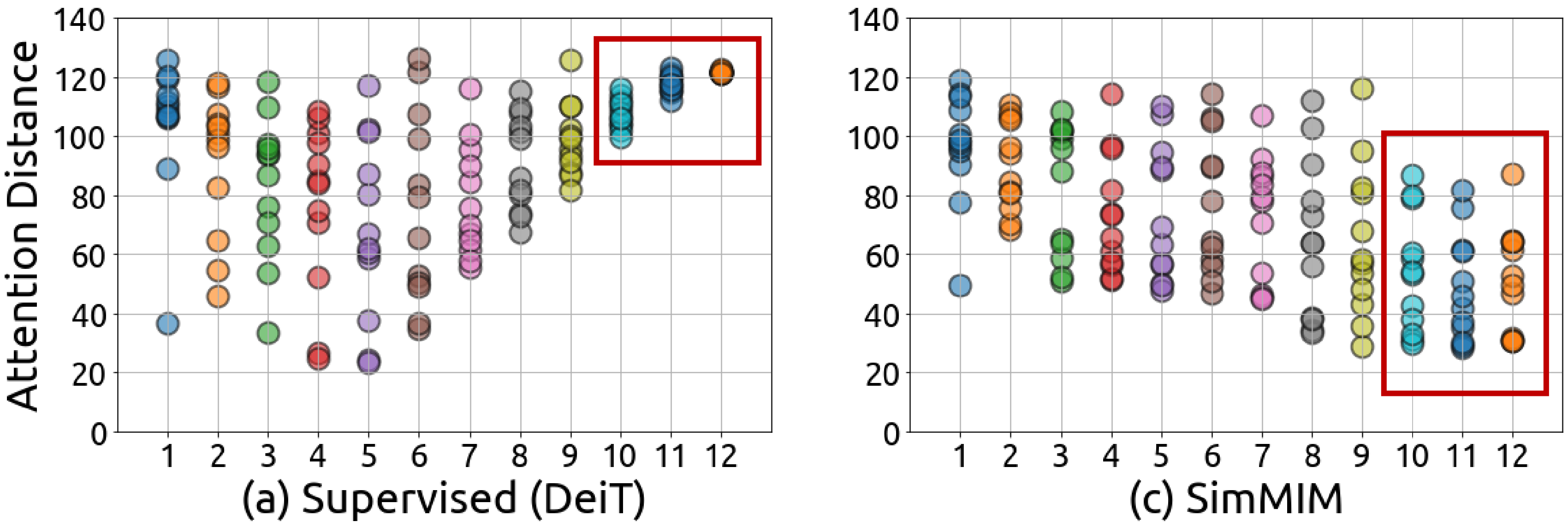
- Could MIM be simple but effective?
  - SimMIM: A Simple Framework for Masked Image Modeling, CVPR 2022
- **How and where does MIM work?**
  - **Revealing the Dark Secrets of Masked Image Modeling, Arxiv 2022**
- Could MIM benefit from larger-scale data?
  - On Data Scaling in Masked Image Modeling, Arxiv 2022

Understanding MIM

- **Visualizations**
  - Local attention or global attention?
  - Diverse attention heads or not?
  - Do features are different across layers?
- Experiments
  - Semantic understanding tasks
  - Geometric and motion tasks
  - Combined tasks

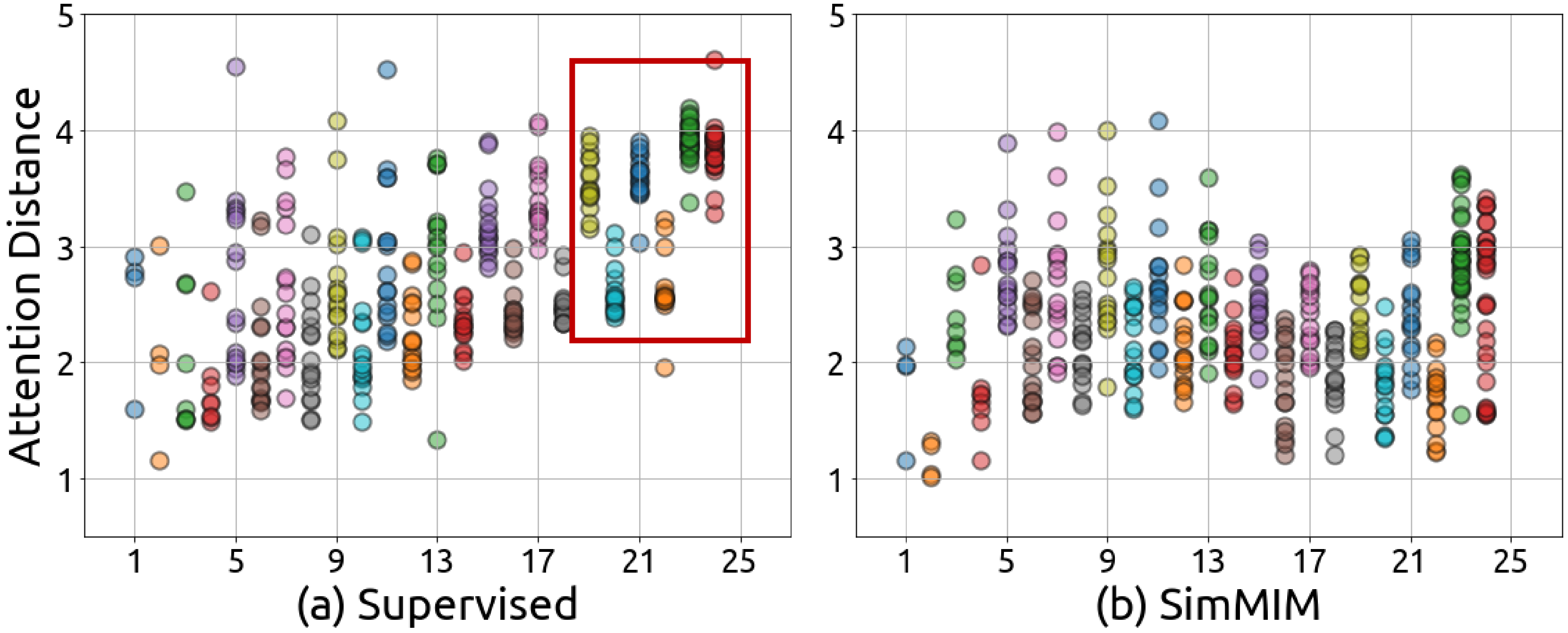
Understanding MIM: A Visualization Perspective

Local Attention v.s. Global Attention (ViT-B)



Understanding MIM: A Visualization Perspective

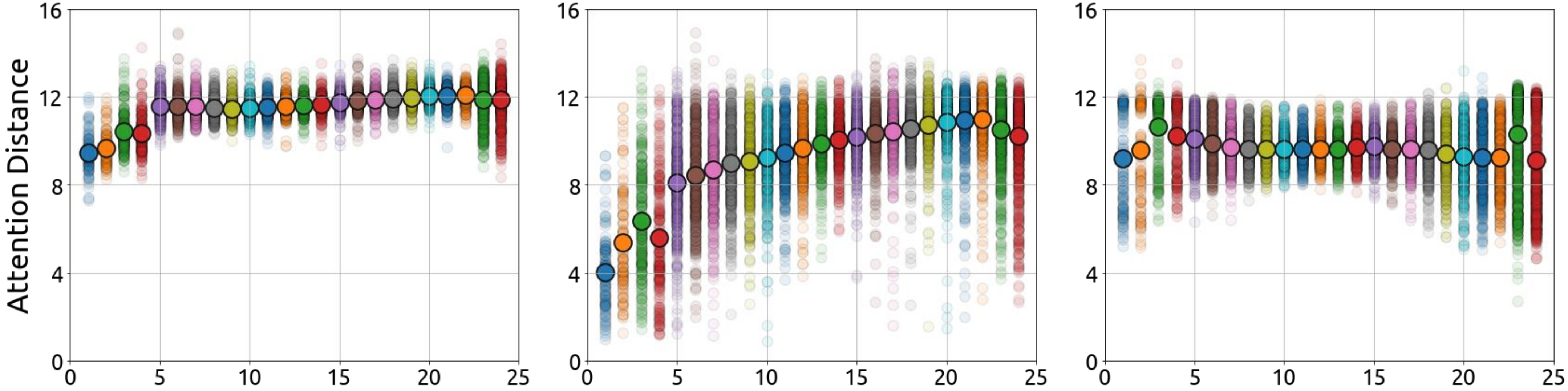
Local Attention v.s. Global Attention (Swin-B)



Understanding MIM: A Visualization Perspective

Could MIM benefit large-kernel ConvNets?

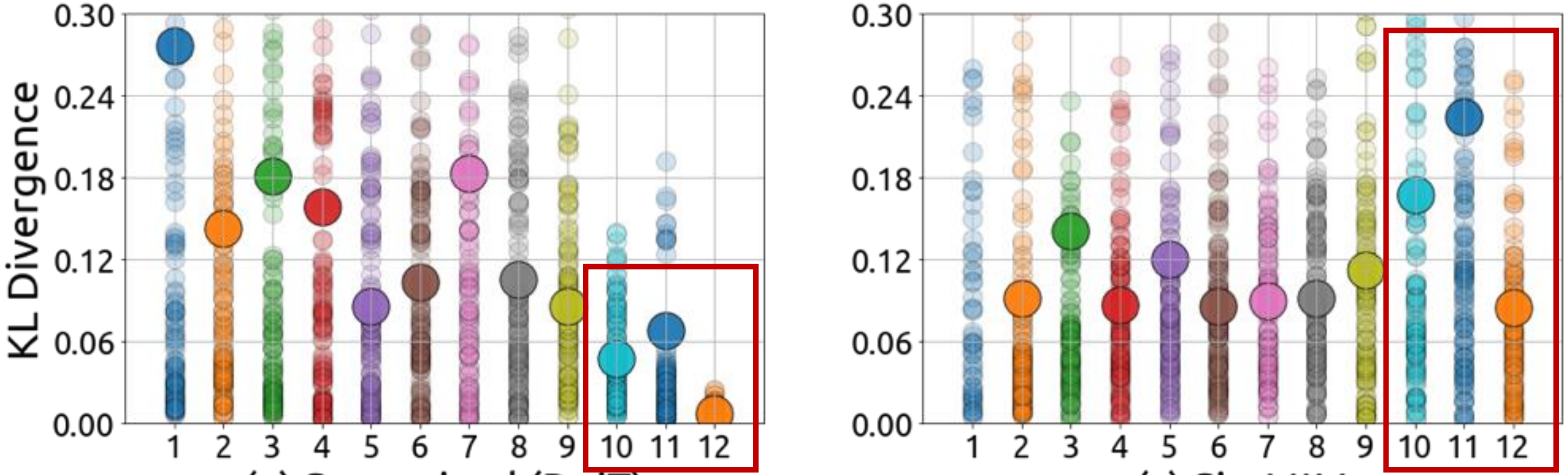
backbone	pre-train	ImageNet-1K	Pose Estimation COCO *val*	Pose Estimation COCO *test*	Pose Estimation Crowd-Pose
RepLKNet-31B	1K-SUP w/ Reparam.	83.5	74.6	73.9	70.2
RepLKNet-31B	1K-MIM w/o Reparam.	83.3	76.5 +1.9	75.8	72.4



(a) Supervised w/o Rep (b) Supervised w/ Rep (c) SimMIM

Understanding MIM: A Visualization Perspective

Diverse attention heads or not? (ViT-B)

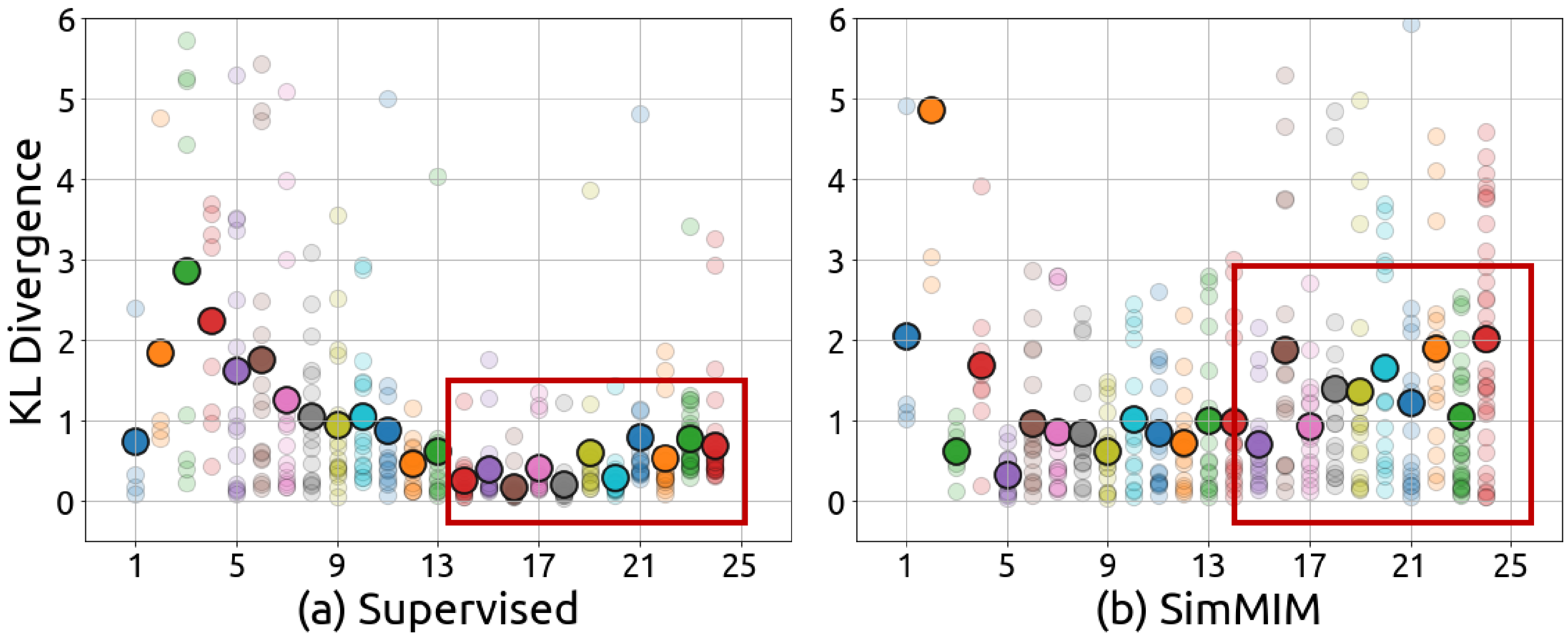


(a) Supervised (DeiT)

(c) SimMIM

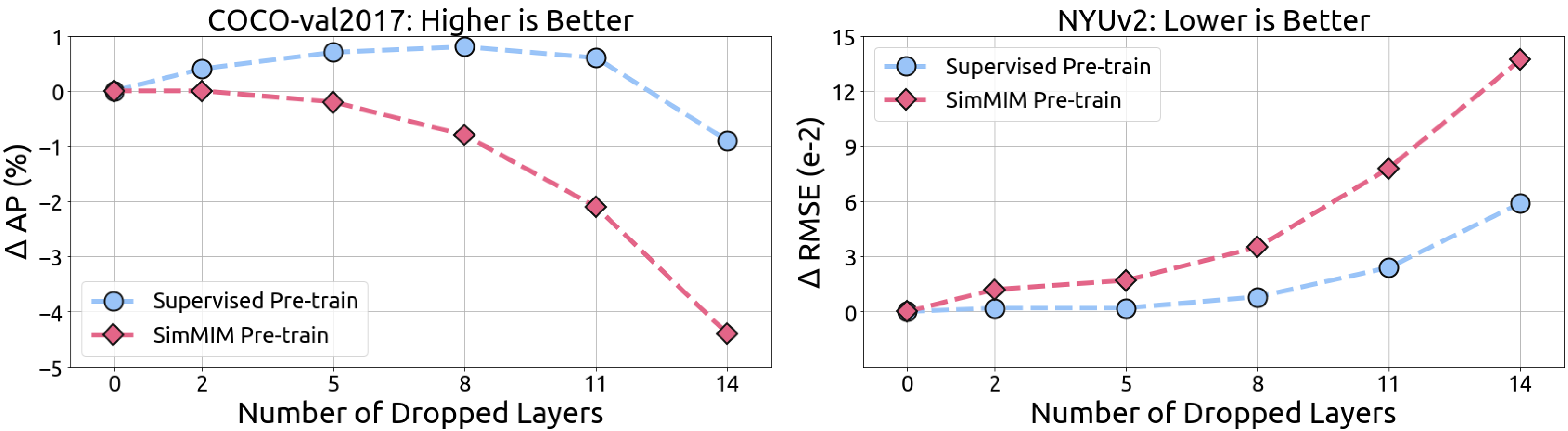
Understanding MIM: A Visualization Perspective

Diverse attention heads or not? (Swin-B)



Understanding MIM: A Visualization Perspective

Less diversity on attention heads would harm the downstream performance.

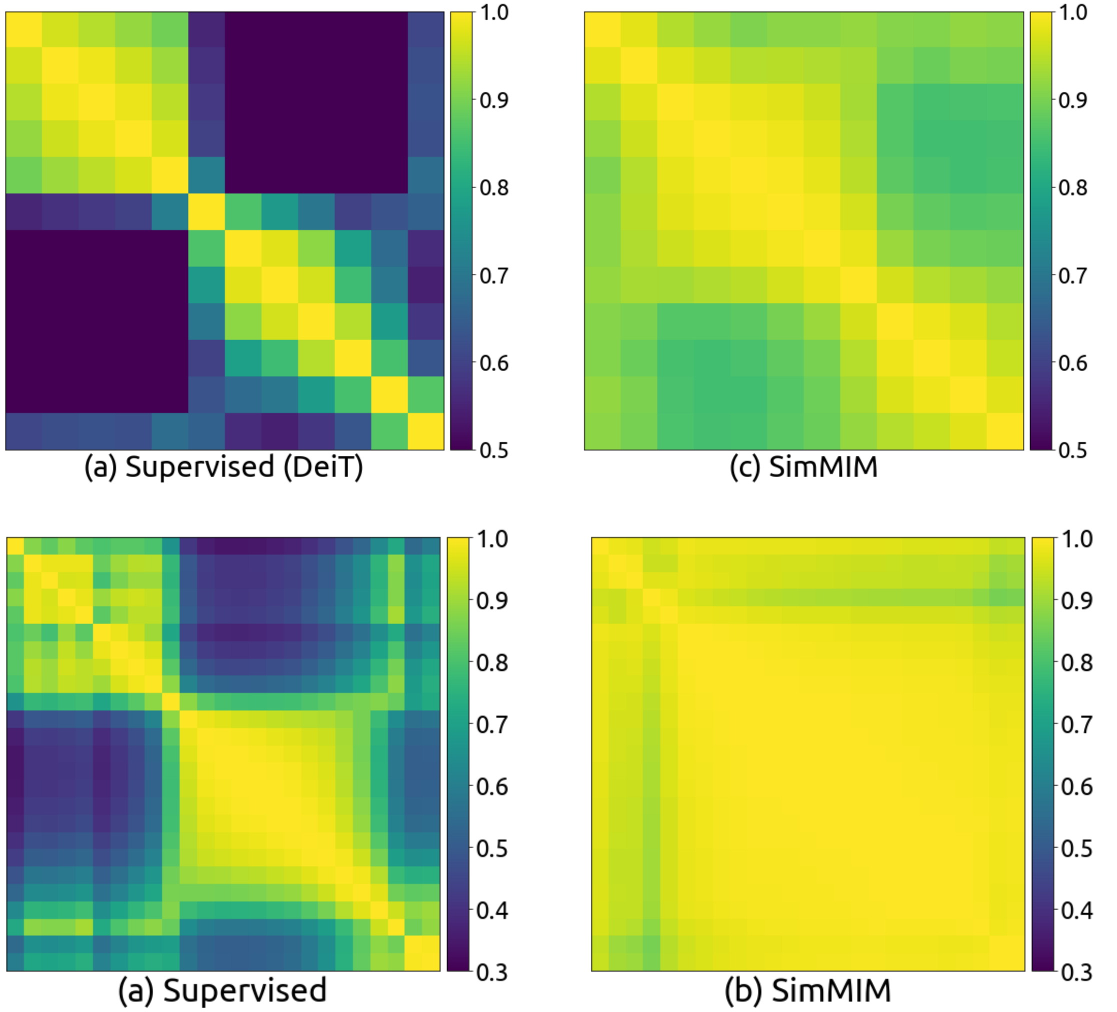


Understanding MIM: A Visualization Perspective

The difference on features across layers (via CKA)

ViT-B

Swin-B



Understanding MIM

- Visualizations
    - Local attention or global attention?
    - Diverse attention heads or not?
    - Do features are different across layers?
- **Experiments**
    - Semantic understanding tasks
    - Geometric and motion tasks
    - Combined tasks

Understanding MIM : An Experimental Perspective

Semantic Understanding Tasks

pre-train	Concept Generalization (CoG)					Kornlith et al's 12 datasets (K12)					iNat18
	L_1	L_2	L_3	L_4	L_5	Food	Birdsnap	Cars	Aircraft	Average (7)	
1K-SUP	79.4	76.2	72.7	72.5	68.4	93.2	81.8	88.6	83.0	89.7	77.7
1K-MIM	79.6	77.1	73.6	73.0	69.1	94.2	83.7	89.2	83.5	86.1	79.6

Methods	Food101	Birdsnap	Stanford Cars	FGVC Aircraft	Oxford Pets	Caltech101	Flowers102	DTD	SUN397	CIFAR10	CIFAR100
1K-SUP	93.2	81.7	88.6	83.0	95.9	91.9	97.7	80.3	72.3	99.1	91.0
1K-MIM	94.2	83.7	89.2	83.5	90.9	85.5	91.4	73.4	70.8	99.2	91.4

Understanding MIM : An Experimental Perspective

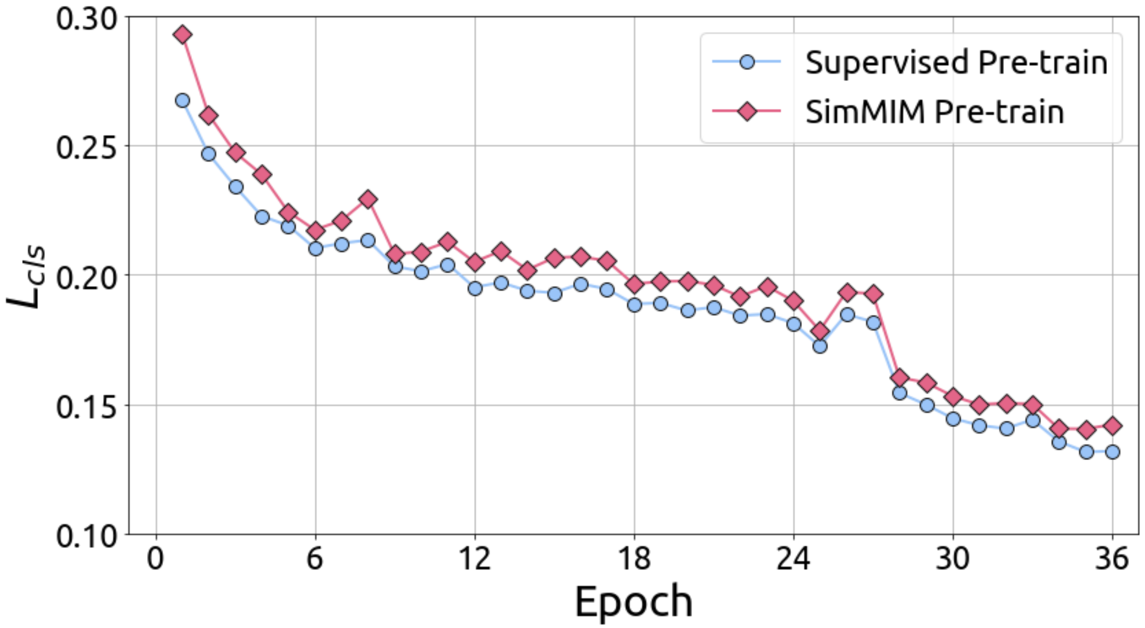
Geometric and Motion Tasks

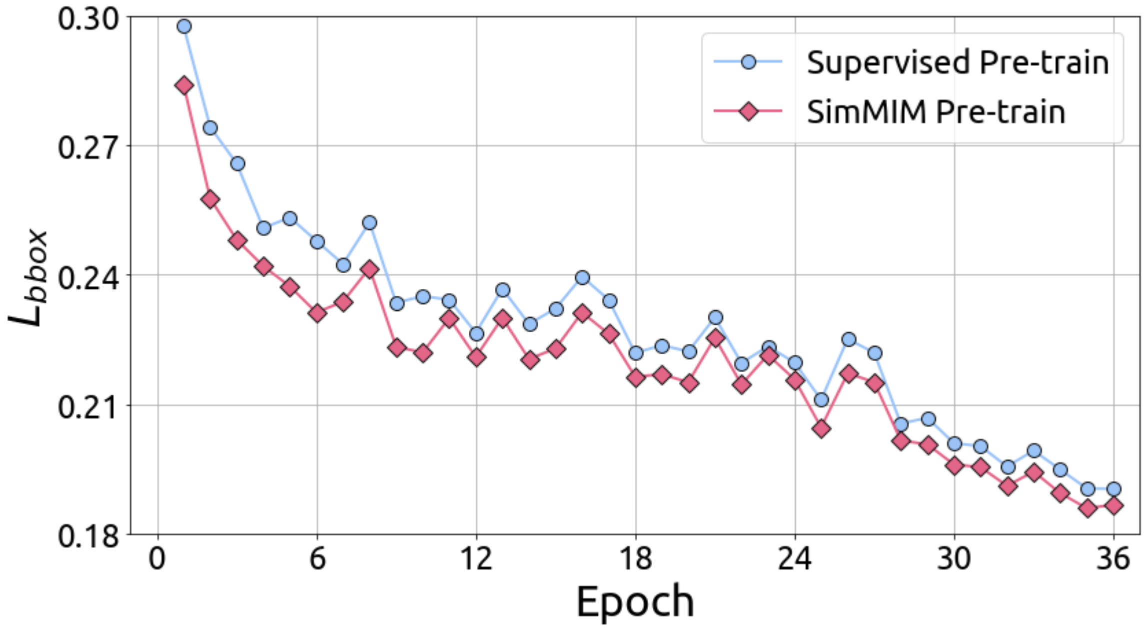
backbone	pre-train	Pose Estimation			Depth Estimation		Video Object Tracking		
		COCO *val*	COCO *test*	Crowd-Pose	NYUv2	KITTI	GOT10k *test*	Track-Net	LaSOT
SwinV2-B	1K-SUP	75.2	74.5	70.7	0.352	2.313	70.1	81.5	69.4
	22K-SUP	75.9	75.1	72.2	0.335	2.240	69.9	81.0	67.8
	1K-MIM	**77.6** +2.4	**76.7**	**74.9**	**0.304**	**2.050**	**70.8**	**82.0**	**70.0**
SwinV2-L	22K-SUP	76.5	75.7	72.7	0.334	2.150	71.1	81.5	69.2
	1K-MIM	**78.1** +0.9	**77.2**	**75.5**	**0.287** +0.043	**1.966**	**72.9**	**82.5**	**70.7** +0.6
Representative methods		HRFormer [79]			BinsFormer [50]		MixFormer [12]		
		77.2	76.2	72.5	0.330	2.098	75.6	83.9	70.1

Understanding MIM : An Experimental Perspective

Combined Tasks: object detection & semantic segmentation

backbone	pre-train	Object Det. (COCO) Mask R-CNN AP^{box}	Object Det. (COCO) Mask R-CNN AP^{mask}	Semantic Seg. (ADE-20K) UperNet mIoU	Semantic Seg. (ADE-20K) Mask2former mIoU
SwinV2-B	1K-SUP	$\boxed{51.9}$	45.7	50.9	52.3
	1K-MIM	$\boxed{52.9}$	46.7	49.3 (−1.6)	51.7 (−0.6)





Masked Image Modeling

- Could MIM be simple but effective?
  - SimMIM: A Simple Framework for Masked Image Modeling, CVPR 2022
- How and where does MIM work?
  - Revealing the Dark Secrets of Masked Image Modeling, Arxiv 2022
- **Could MIM benefit from larger-scale data?**
  - **On Data Scaling in Masked Image Modeling, Arxiv 2022**

Data Scaling of MIM: Setup

Model	Base Channel	Depth	Head	Window Size		Backbone Params
				pre-train	fine-tune	
SwinV2-S	96	{2, 2, 18, 2}	{3, 6, 12, 24}	12	14	49M
SwinV2-B	128	{2, 2, 18, 2}	{4, 8, 16, 32}	12	14	87M
SwinV2-L	192	{2, 2, 18, 2}	{6, 12, 24, 48}	12	14	195M
SwinV2-H	352	{2, 2, 18, 2}	{11, 22, 44, 88}	12	14	655M
SwinV2-G	448	{2, 2, 18, 2}	{14, 28, 56, 112}	12	14	1061M

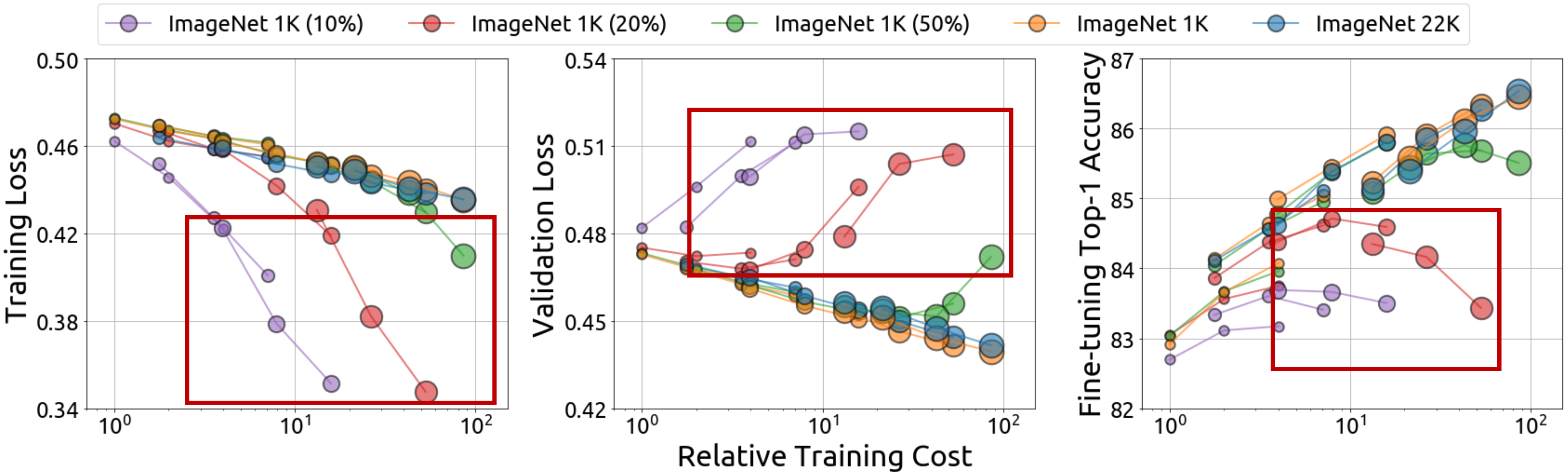
Architecture Specifications

	IN1K (10%)	IN1K (20%)	IN1K (50%)	IN100	IN1K(100%)	IN22K(100%)
#Classes	1×10^3	1×10^3	1×10^3	1×10^2	1×10^3	2.18×10^4
#Images	1.28×10^5	2.56×10^5	6.41×10^5	1.27×10^5	1.28×10^6	1.42×10^7

Dataset Specifications

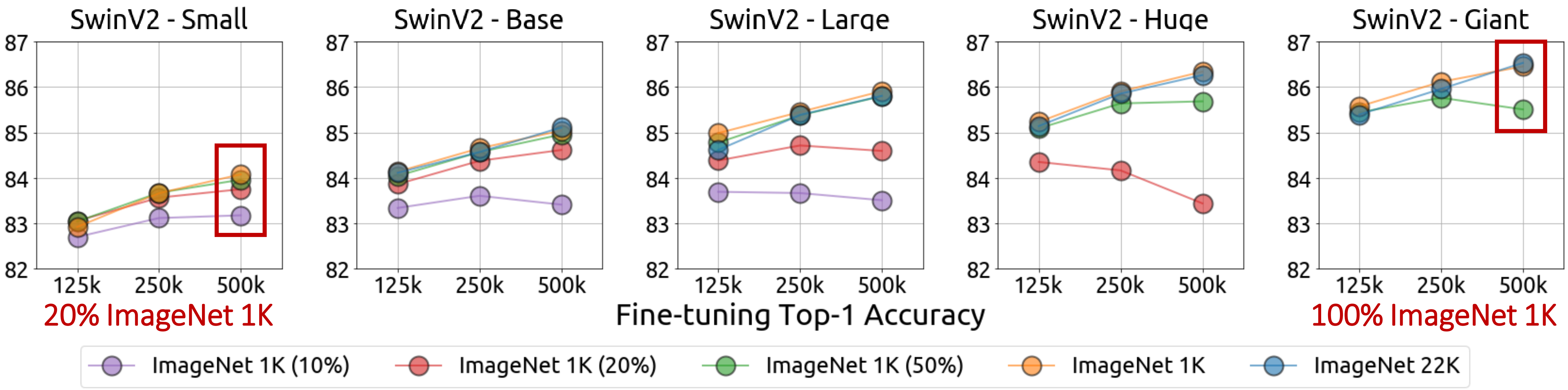
Data Scaling of MIM: Experiments

Masked image modeling remains **demanding for large datasets**



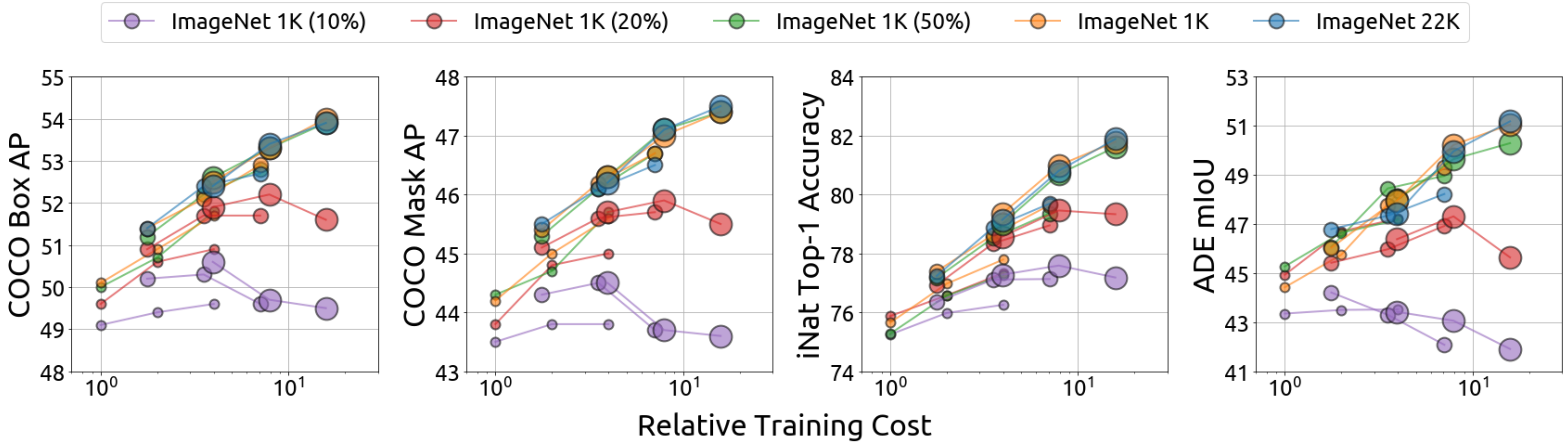
Data Scaling of MIM: Experiments

- **Training length matters**: less overfitting with 125k iterations
- Large models can benefit from more data at a **longer training length**



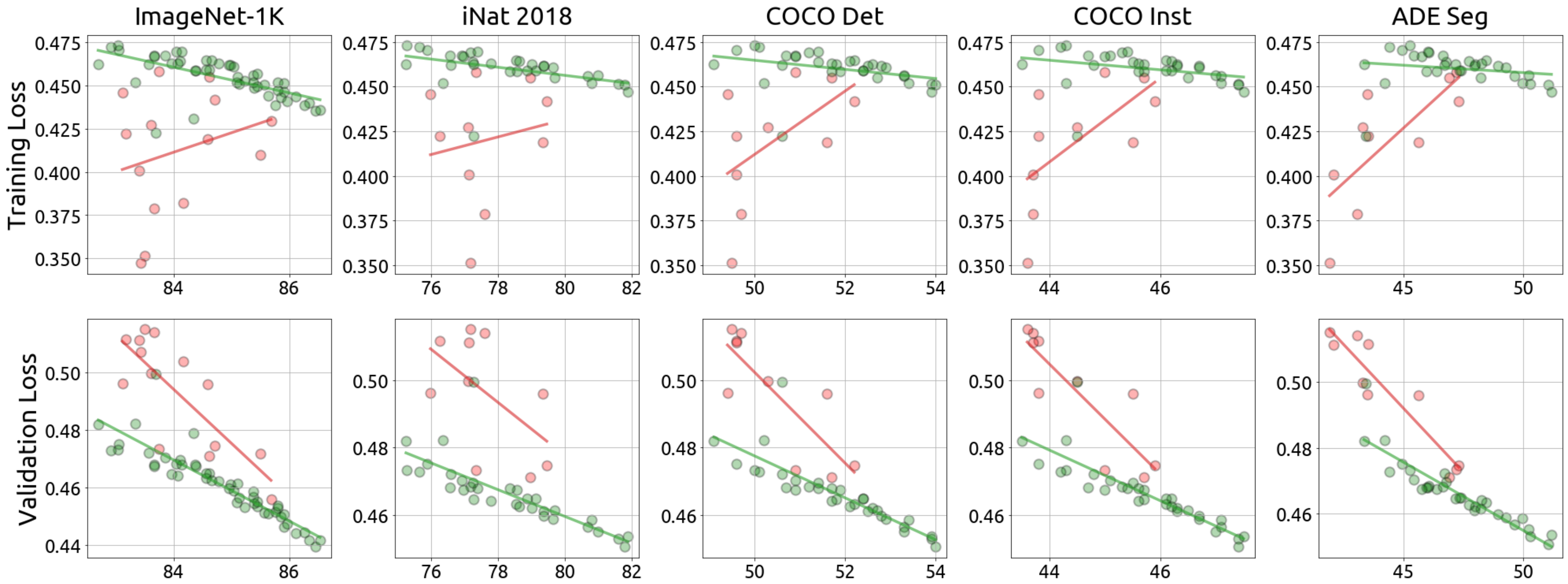
Data Scaling of MIM: Experiments

- This observation is kept across tasks
  - COCO Object Det. & iNaturalist 2018 Cls. & ADE-20K Semantic Seg.



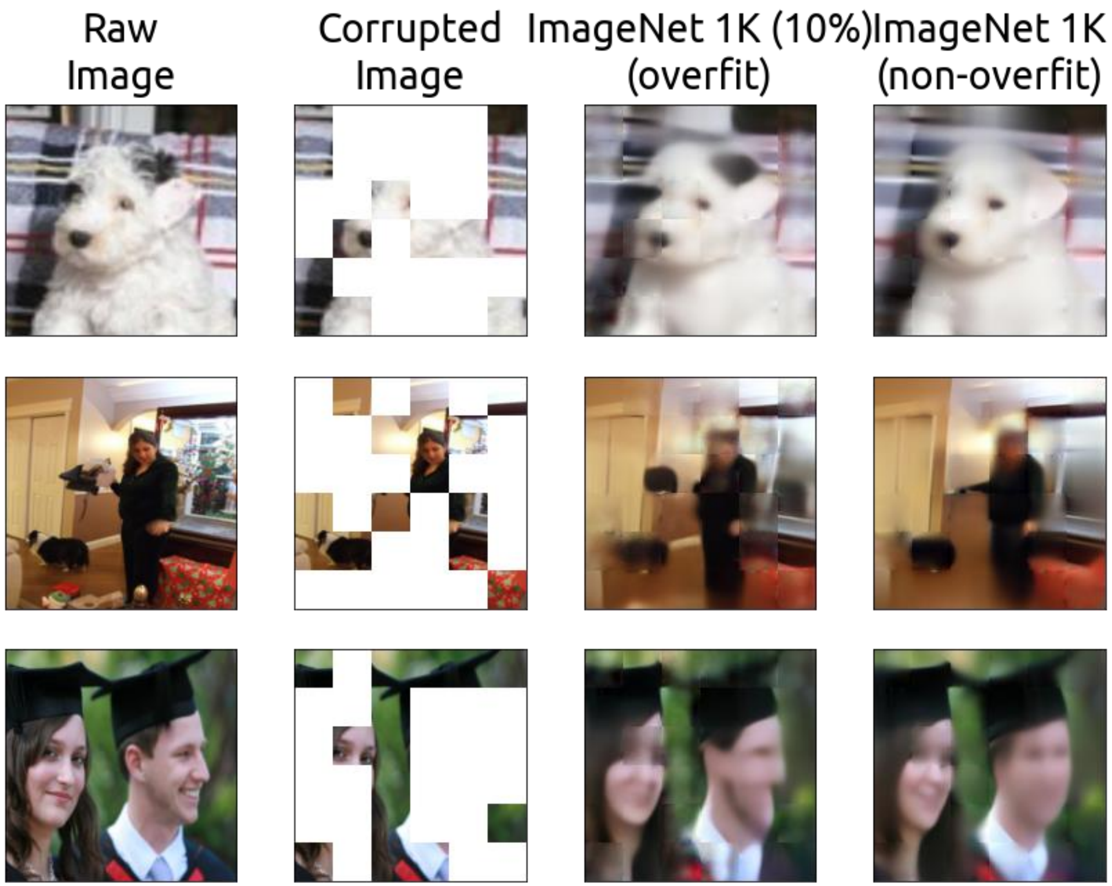
Data Scaling of MIM: Correlation Analysis

The validation loss is a good **proxy metric** of the fine-tuning performance

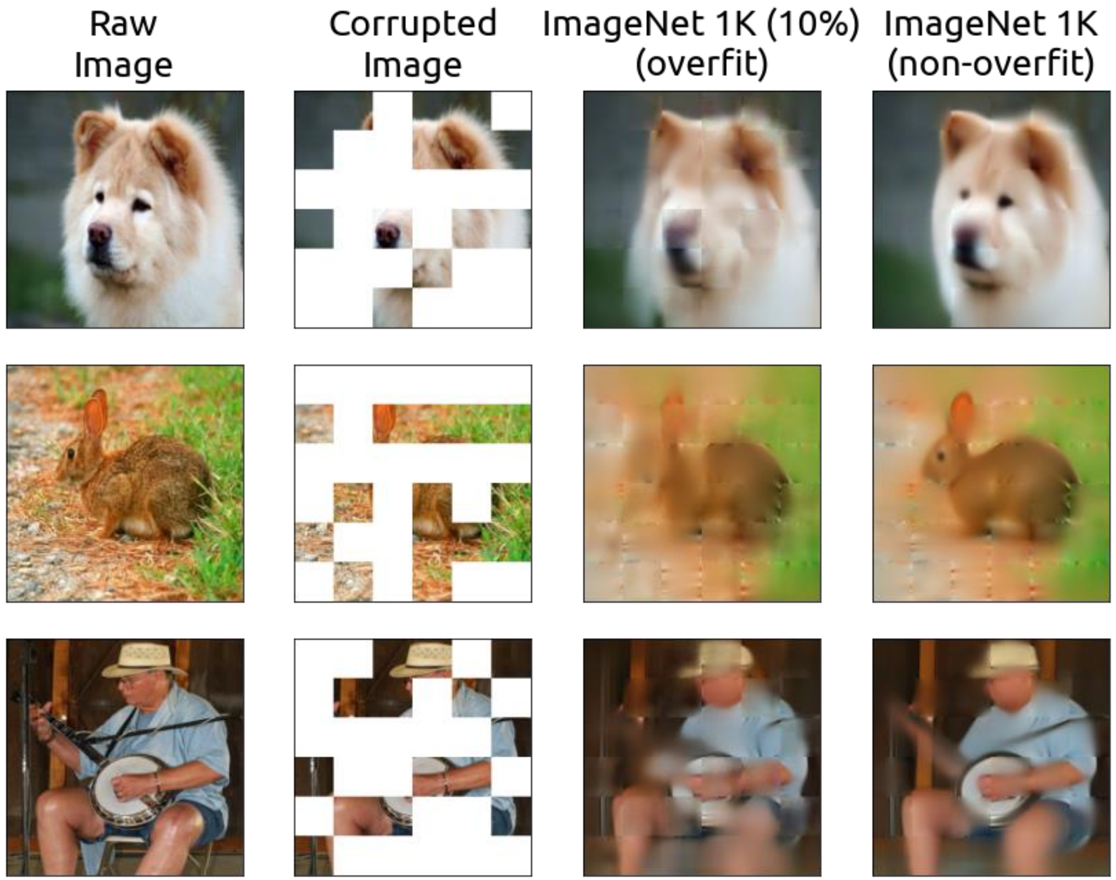


(a) Top-1 Acc (b) Top-1 Acc (c) bbox mAP (d) mask mAP (e) mean IoU

Data Scaling of MIM: Visualizations



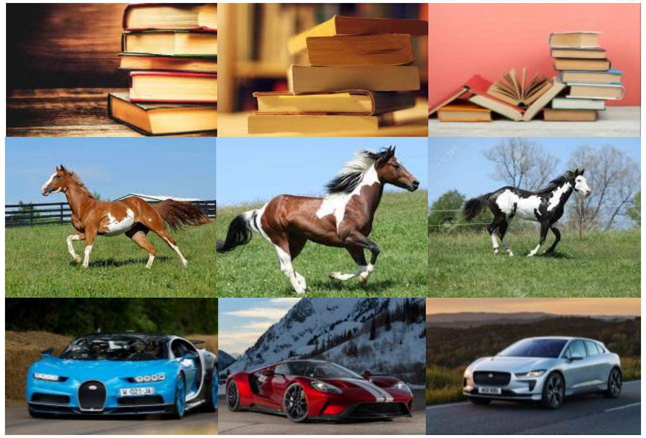
Training Set

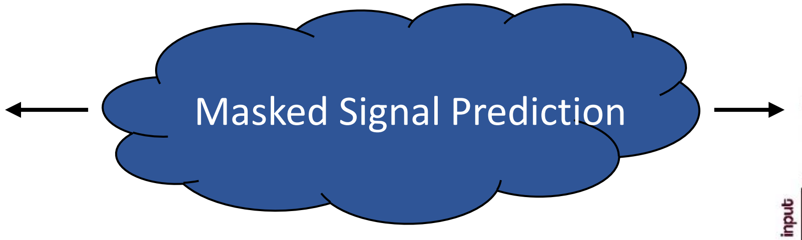


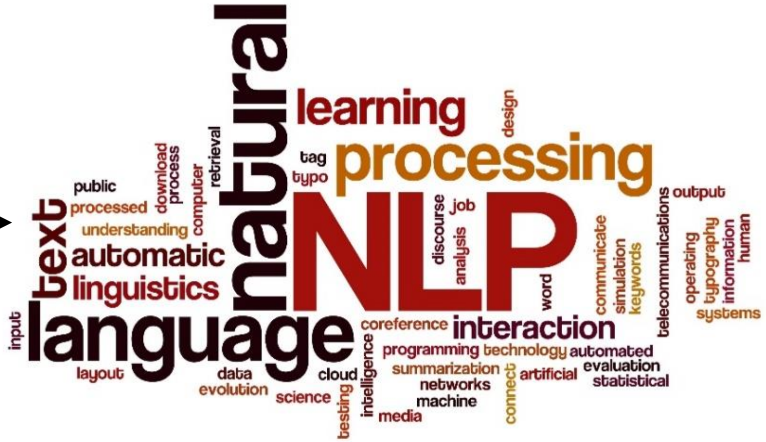
Validation Set

Takeaway

- Approach: A Simple but effective MIM framework (SimMIM)
- Understanding: How and where MIM works
- Data scaling: MIM could still benefit from larger dataset
- -> Task convergence between CV and NLP







Thanks All!
Q & A