NANYANG TECHNOLOGICAL UNIVERSITY

MMLab @NTU

Code Link

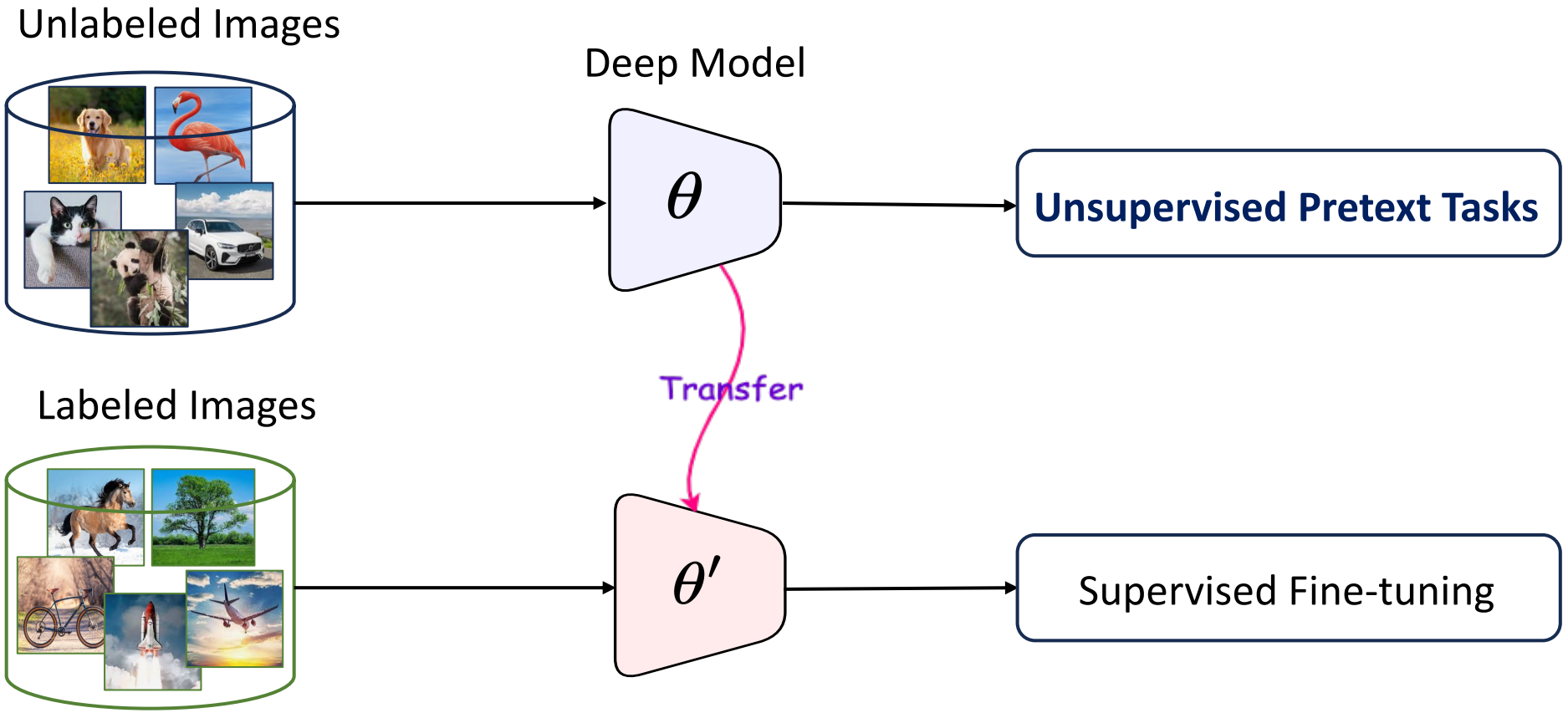
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Correlational Image Modeling for Self-Supervised Visual Pre-Training

Wei Li, Jiahao Xie, Chen Change Loy

Background

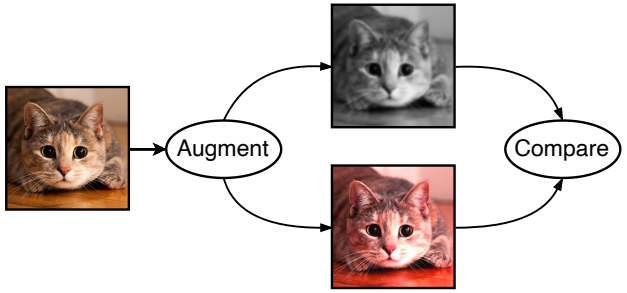
- Self-Supervised Visual Pre-training



Background

- Pretext Tasks in SSL

Multi-View Self-Supervised Learning



(a) MV-SSL

augment-and-compare

[MoCo He et.al],
[DINO Caron et.al],
[BYOL Grill et.al],
.....

Masked Image Modeling

(b) MIM

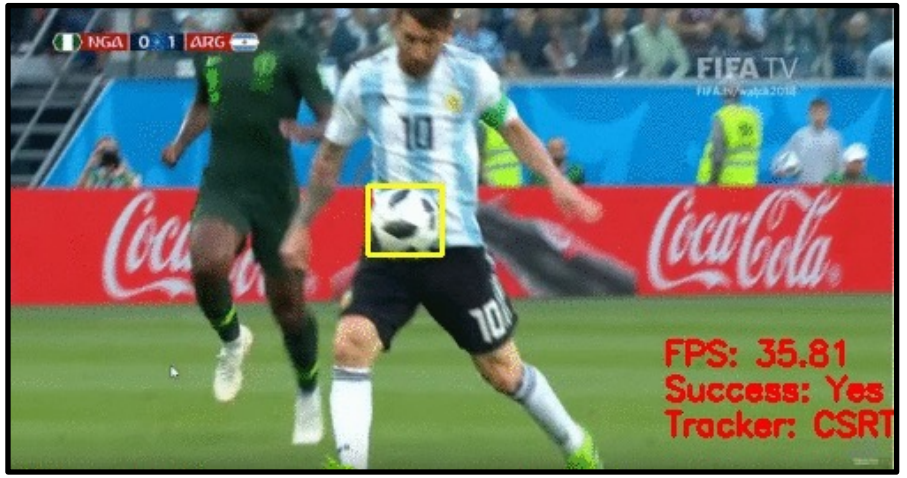
mask-and-predict

[MAE He et.al],
[MFM Xie et.al],
[BEiT Bao et.al],
.....

?

Motivations

- Visual Tracking



Visual Tracking Benchmarks: https://www3.cs.stonybrook.edu/~hling/VTB/index.html

Nanonets: https://nanonets.com/blog/object-tracking-deepsort/

Motivations

- Deep Visual Tracking

target or exemplar

context image

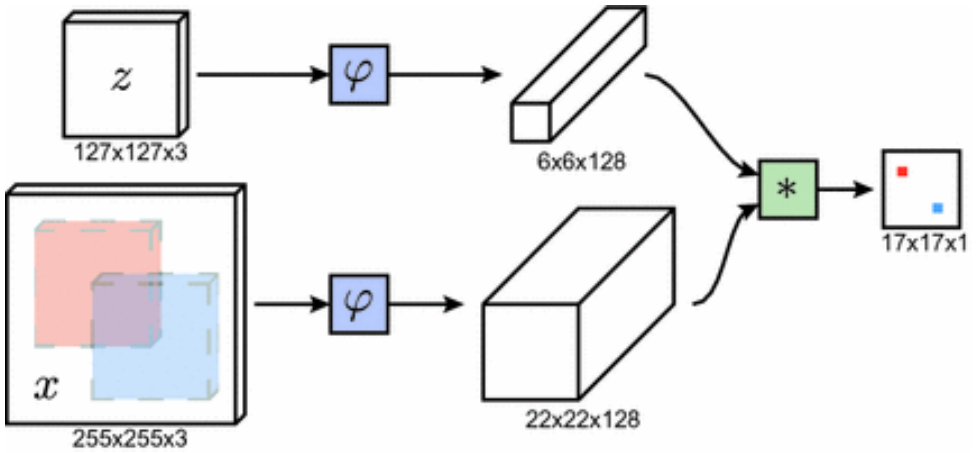
$$f(\boldsymbol{z}, \boldsymbol{c}) = f_\theta(\boldsymbol{z}) \star f_\theta(\boldsymbol{c}) + \boldsymbol{b}\mathbb{1}$$

backbone model

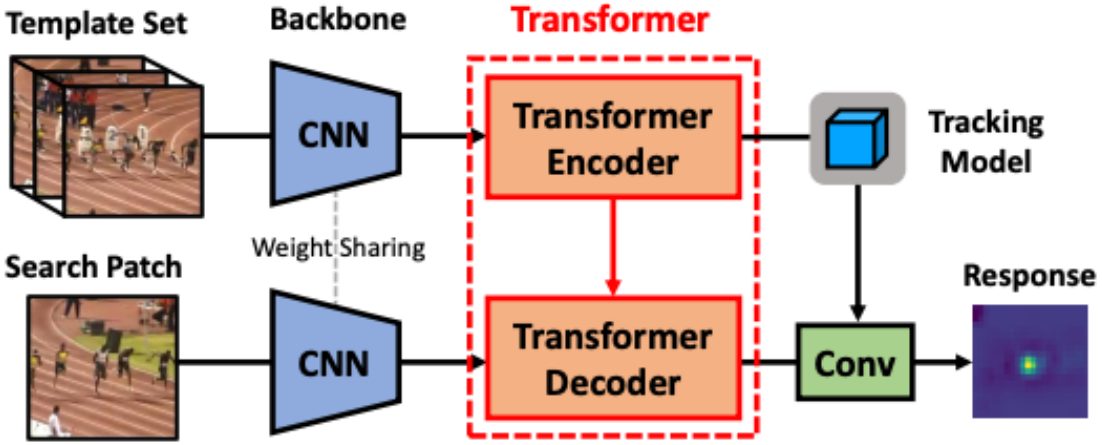
correlation operator

Motivations

- Deep Visual Tracking



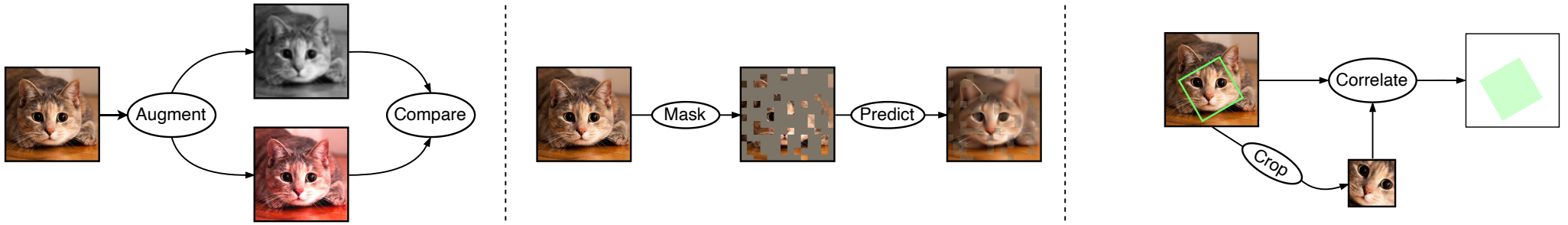
Bertinetto et.al, ECCV-W 2016



Wang et.al, CVPR 2021

Approach

- Correlational Image Modeling



(a) MV-SSL

(b) MIM

(c) CIM

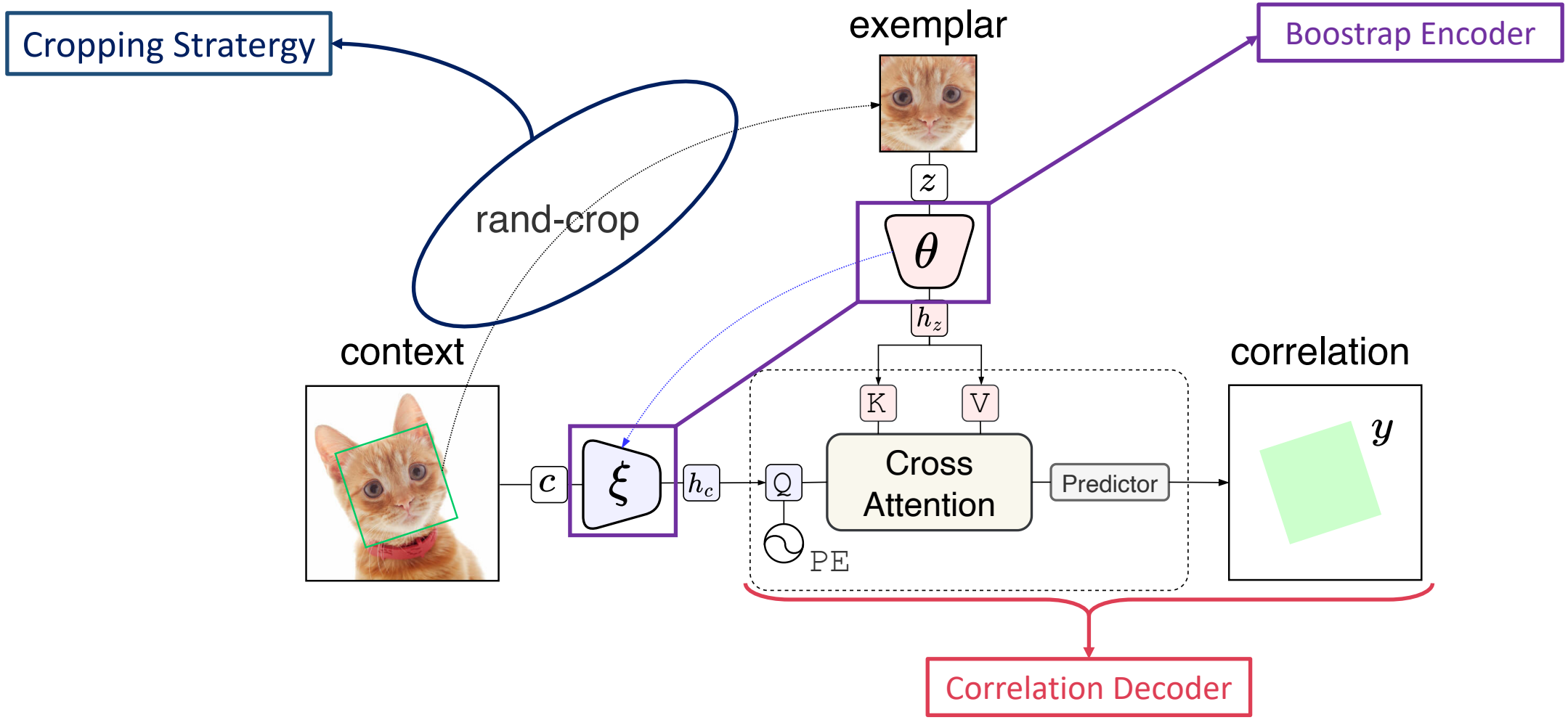
augment-and-compare

mask-and-predict

crop-and-correlate

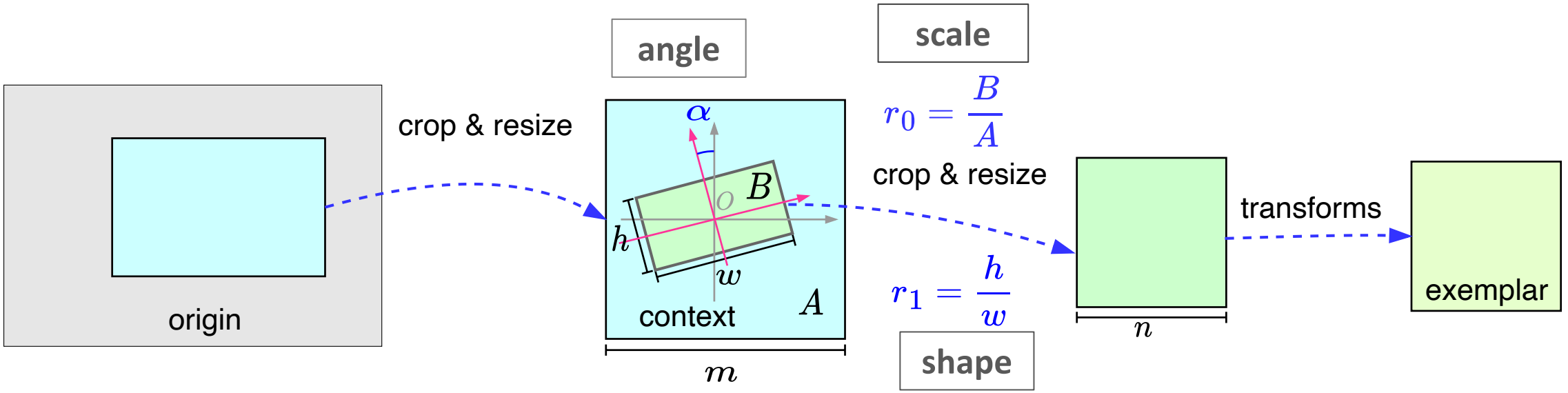
Approach

- Correlational Image Modeling



Approach

- Correlational Image Modeling



Experiments

- ImageNet-1K Image Classification

Method	Pre-train Data	Pretext Task	Tokenizer	Epochs	ViT-S	ViT-B
Scratch [45]	-	-	-	-	79.9	81.8
MP3 [71]	IN-1K	Jigsaw	-	100	-	81.9
MoCo v3 [12]	IN-1K	MV-SSL	-	1200	81.4	83.2
DINO [8]	IN-1K	MV-SSL	-	1600	81.5	82.8
BEiT [2]	IN-1K+DALL-E	MIM	dVAE	300	81.3	82.9
SimMIM [68]†	IN-1K	MIM	-	300	80.9	82.9
MAE [22]†	IN-1K	MIM	-	300	80.6	82.9
CIM	IN-1K	CIM	-	300	81.6	83.1

(a) Vision Transformers

Method	Pretext Task	Epochs	Top-1 acc (%)
Fine-tuning for 100 epochs			
RSB A3 [56]	-	-	78.1
SimMIM [68]†	MIM	300	77.7
CIM	CIM	300	78.6
Fine-tuning for 300 epochs			
RSB A2 [56]	-	-	79.8
SimSiam [11]	MV-SSL	400	79.1
MoCo v2 [10]	MV-SSL	400	79.6
SimCLR [9]	MV-SSL	800	79.9
BYOL [21]	MV-SSL	400	80.0
SwAV [7]	MV-SSL	600	80.1
SimMIM [68]†	MIM	300	79.5
CIM	CIM	300	80.1

(b) CNNs

Experiments

- ADE20K Semantic Segmentation

Method	Pre-train Data	Pretext Task	mIoU (%)
Supervised [45]	IN-1K w/ labels	-	45.3
MoCo v3 [12]	IN-1K	MV-SSL	47.2
DINO [8]	IN-1K	MV-SSL	46.8
BEiT [2]	IN-1K+DALL-E	MIM	47.7
MAE [22]	IN-1K	MIM	48.1
CIM	IN-1K	CIM	48.1

- Robustness Evaluation

Method	Robustness Benchmarks						Orig.
	FGSM	PGD	IN-C (↓)	IN-A	IN-R	IN-SK	
ViT-B/16 model results							
Scratch [56]	46.3	21.2	**48.5**	28.1	44.7	32.0	81.8
MAE [22]	38.9	11.2	52.3	31.5	48.3	33.8	82.9
CIM	**47.4**	**22.7**	49.3	**30.3**	**48.6**	**35.3**	**83.1**
ResNet-50 model results							
Scratch [56]	**20.2**	**3.4**	77.0	6.6	36.0	25.0	78.1
SimMIM [68]	16.8	2.1	77.0	5.7	34.9	24.2	77.7
CIM	19.4	2.5	**73.5**	**8.5**	**37.4**	**27.2**	**78.6**

Experiments

- Visualization

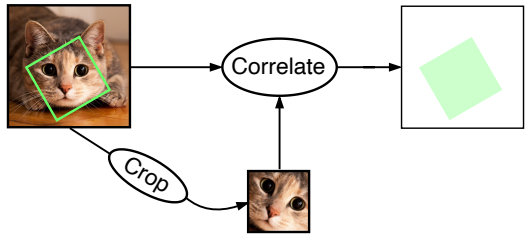
NANYANG TECHNOLOGICAL UNIVERSITY

MMLab @NTU

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Correlational Image Modeling for Self-Supervised Visual Pre-Training

Wei Li, Jiahao Xie, Chen Change Loy



(c) CIM

crop-and-correlate

Code Link