MCMAE: Masked Convolution Meets Masked Autoencoders

Peng Gao[1], Teli Ma[1], Hongsheng Li[12], Ziyi Lin[2], Jifeng Dai[3], Yu Qiao[1]

[1]Shanghai AI Lab [2]MMLab, CUHK [3]Sensetime Research

E-mail: gaopeng@pjlab.org.cn GitHub: https://github.com/Alpha-VL/ConvMAE

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

香 港 中 文 大 學
The Chinese University of Hong Kong

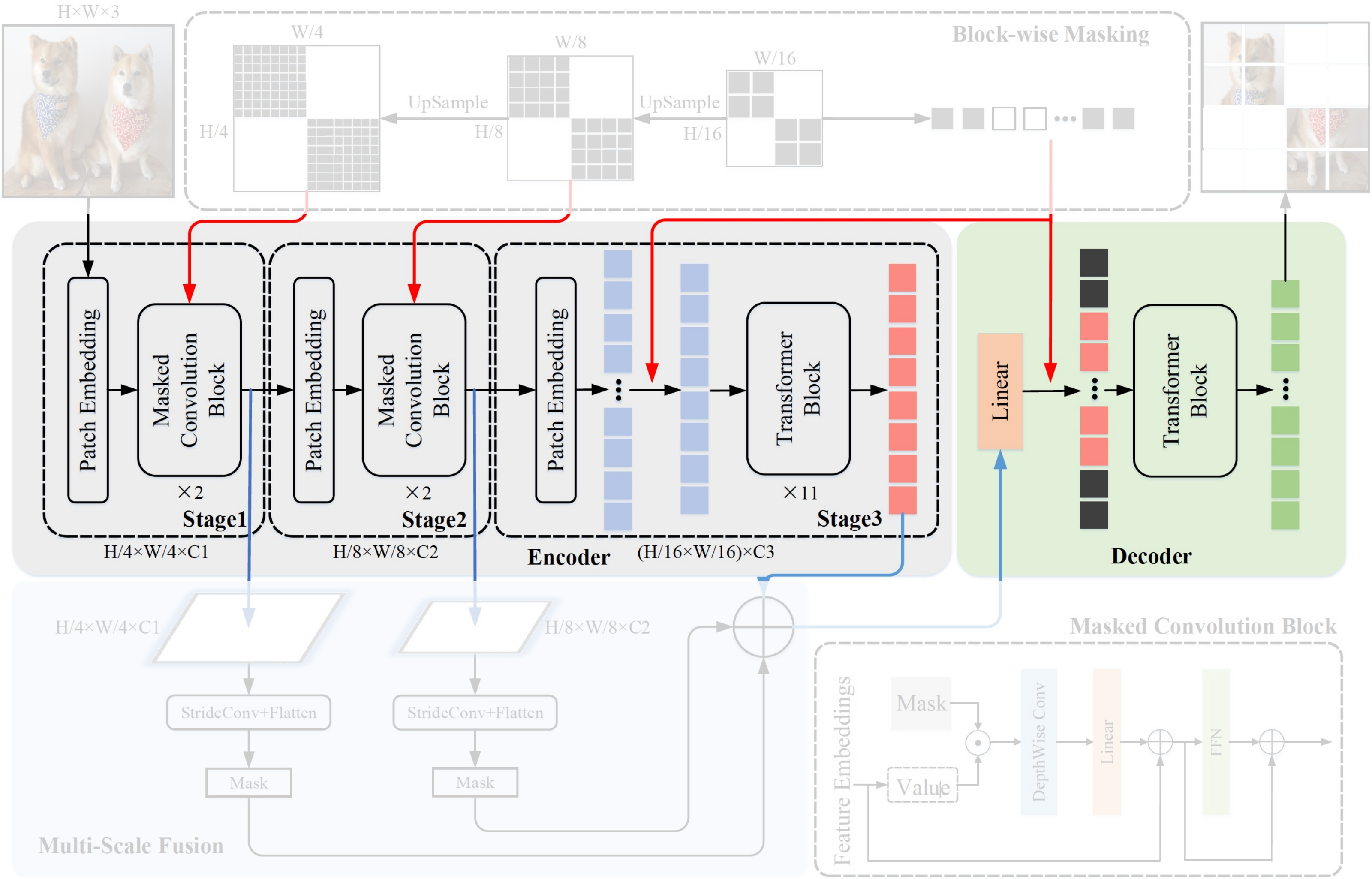
NEURAL INFORMATION PROCESSING SYSTEMS

Background & Motivation

- **Masked Auto-Encoders (MAE)** are effective for pretraining Vision Transformers (ViTs).
- **Convolutions** are effective for supervised learning
- How can we pretrain a mixed Convolution-Transformer architecture, that can take advantage of both
  - the pretraining methods for Transformers and
  - the local inductive bias of convolutions?

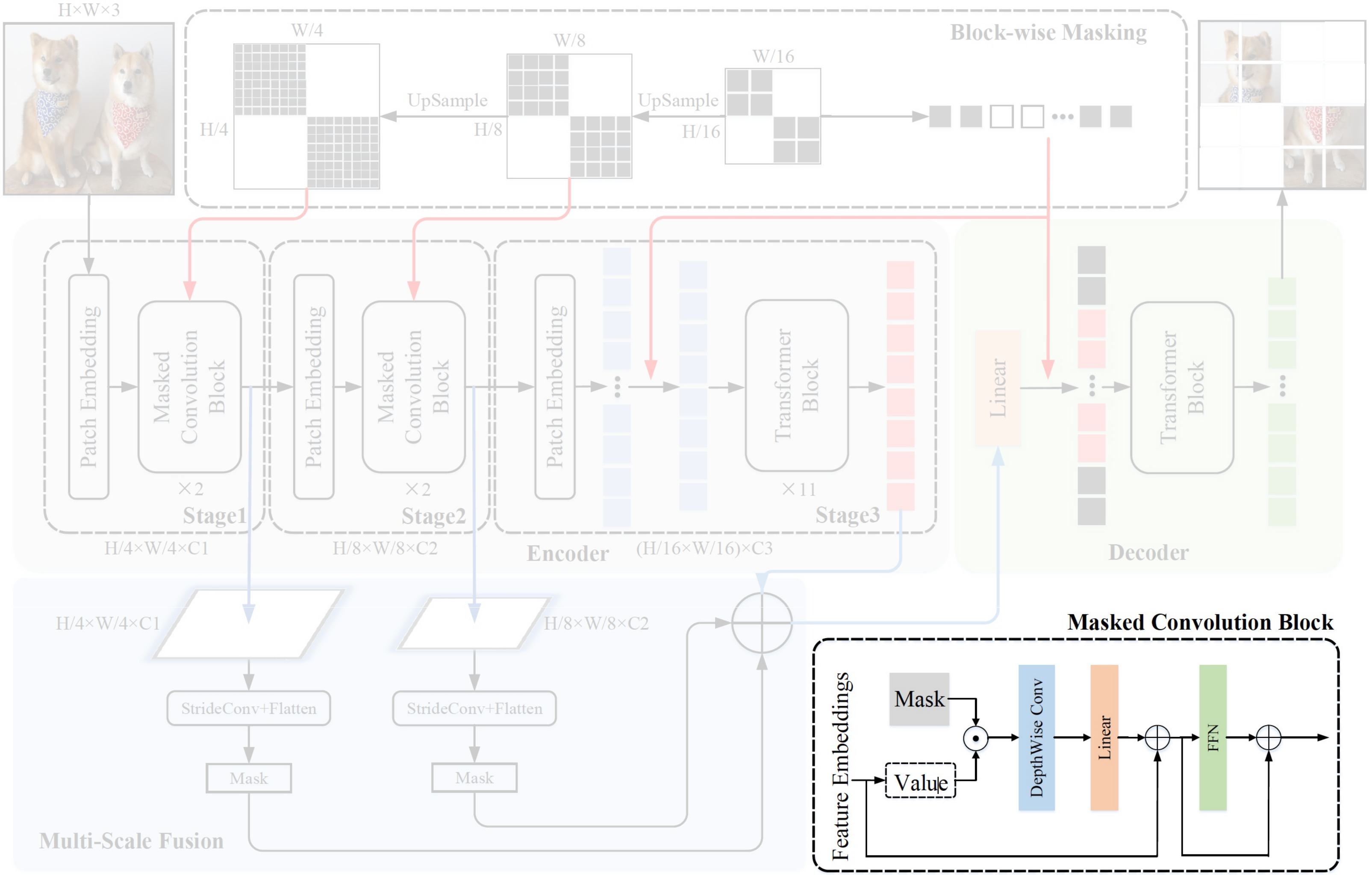
Method – Encoder and Decoder

- Encoder (Early)
  - Convolutional blocks
  - High resolution, local receptive field
- Encoder (Late)
  - Transformer blocks
  - Low resolution, global receptive field
- Decoder follows regular MAE



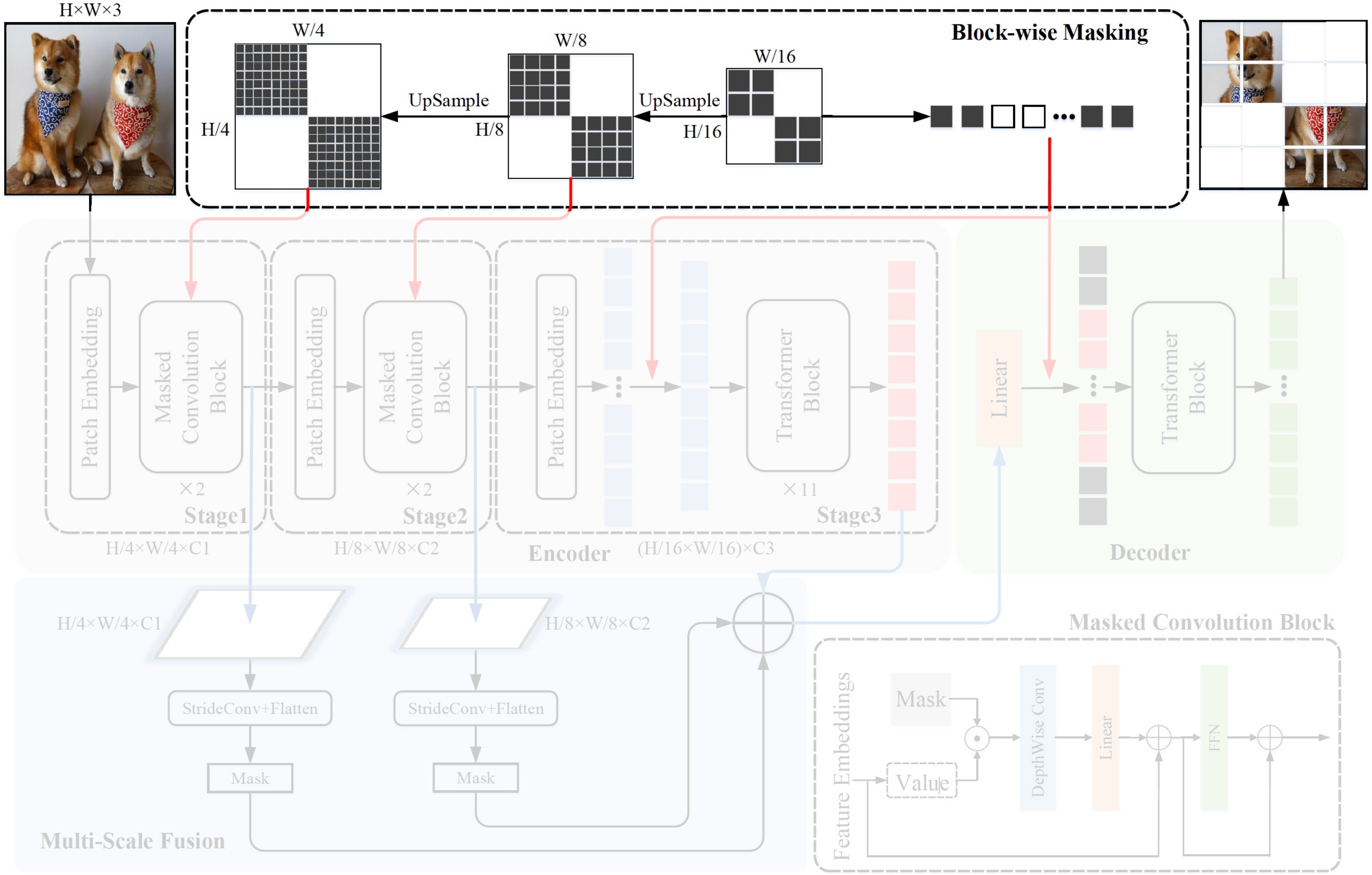
Method – Multi-scale Masking

- ‘Ground-truth leakage’ of convolutions
  - Regular convolutions pass information of adjacent pixels
  - This significantly weakens the pretraining task (reconstructing invisible patches)
- Input to convolutions are masked, forming ‘masked convolutions’ for pretraining.



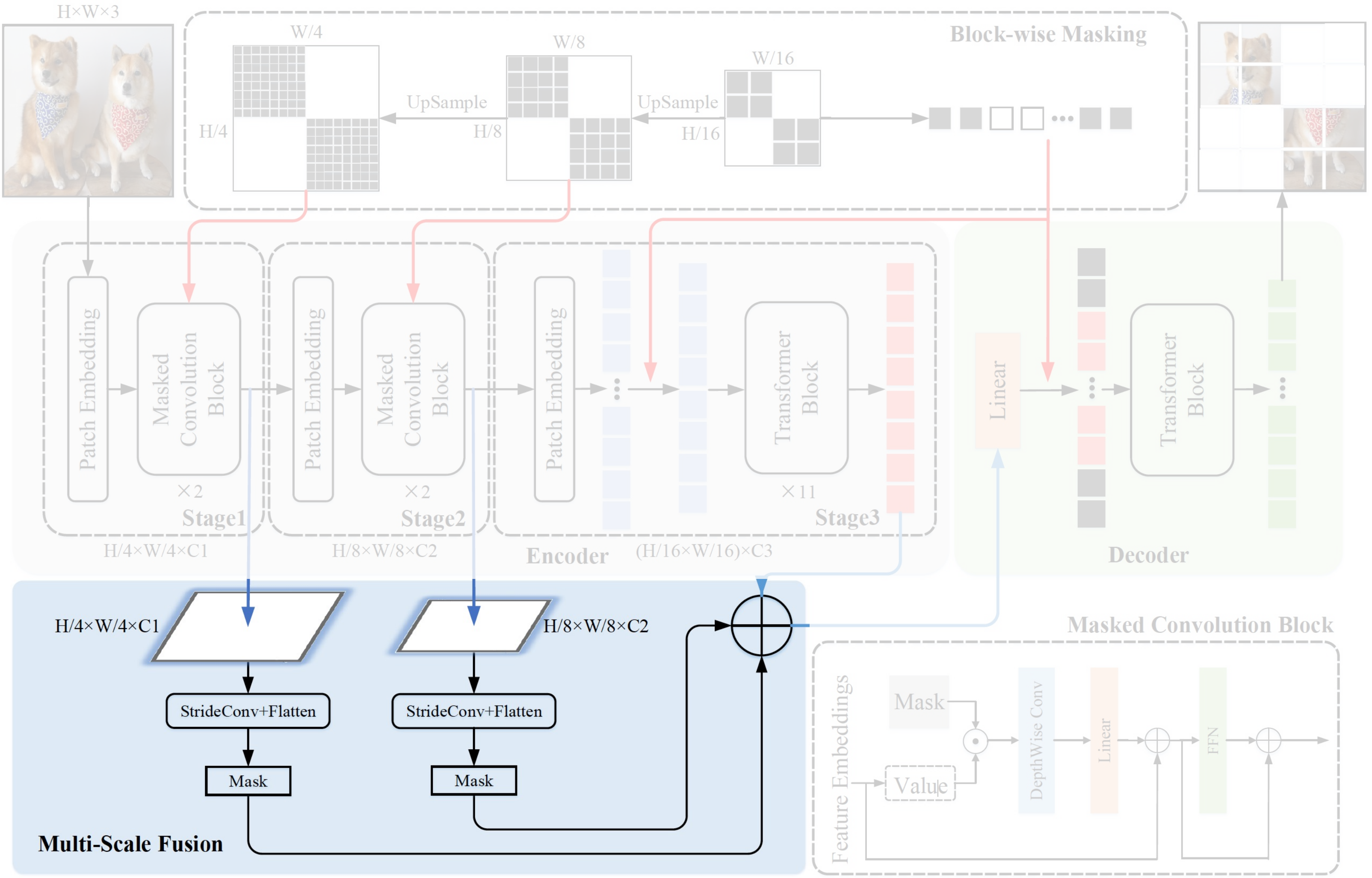
Method – Multi-scale Masking

- We also want sparse visible tokens in the final stage
- Mask at the lowest resolution, up-sample to higher resolutions



Method – Multi-scale Decoding

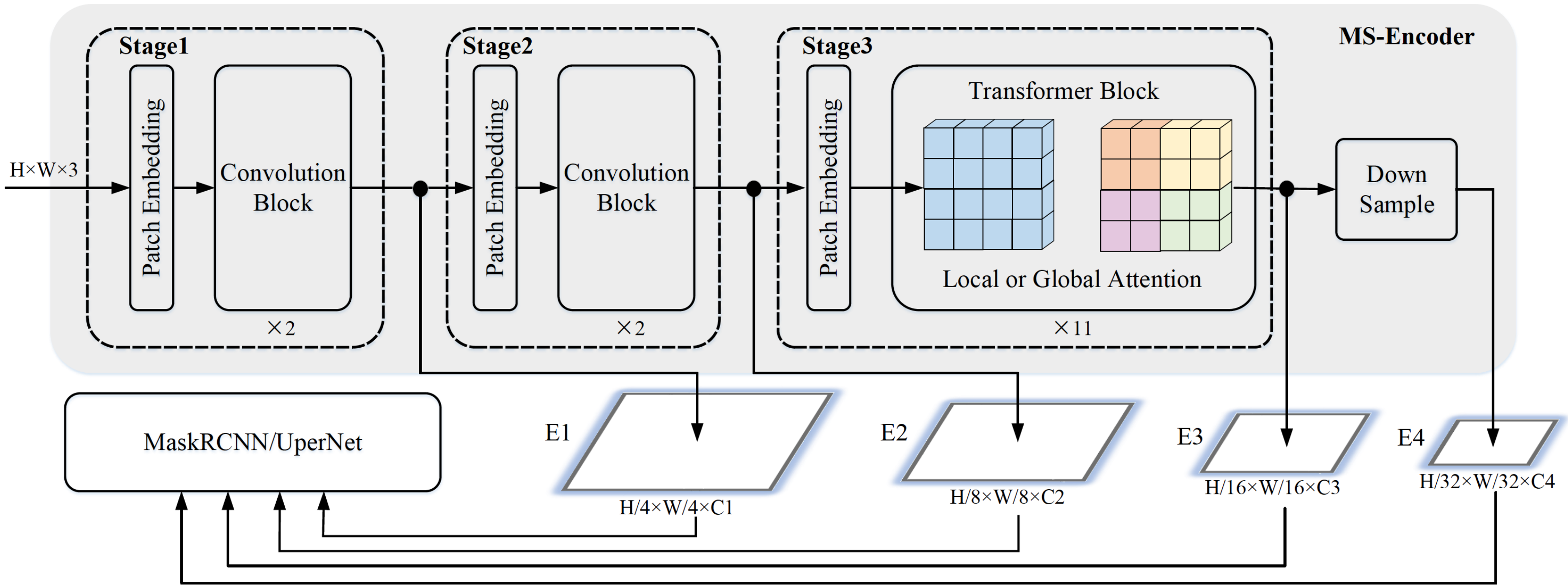
- Outputs of all stages are fed into decoder for pretraining
  - Stronger supervision for earlier stages
  - Bypasses low-level details that are less useful for high-level semantic understanding



Method – MCMAE for Downstream Tasks

NEURAL INFORMATION PROCESSING SYSTEMS

- For object detection & semantic segmentation
  - The multi-scale feature our models produce are especially beneficial



- For video recognition
  - Inputs, positional embeddings, convolutions, etc. changed from 2D to 3D

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments – Main Results on 4 Tasks

Image Classification (ImageNet-1k)

Methods	Backbone	Params. (M)	Supervision	Encoder	P-Epochs	FT (%)	LIN (%)
BEiT [2]	ViT-B	88	DALLE	100%	300	83.0	37.6
MAE [28]	ViT-B	88	RGB	25%	1600	83.6	67.8
SimMIM [59]	Swin-B	88	RGB	100%	800	84.0	N/A
MaskFeat [55]	ViT-B	88	HOG	100%	300	83.6	N/A
data2vec [1]	ViT-B	88	Momentum	100%	800	84.2	N/A
MCMAE	CViT-B	88	RGB	25%	1600	85.0	70.9

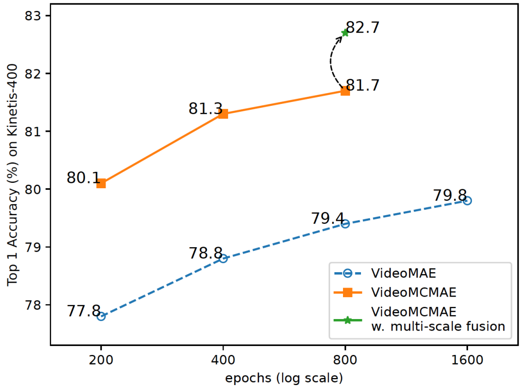
Object Detection with Mask-RCNN (MS-COCO)

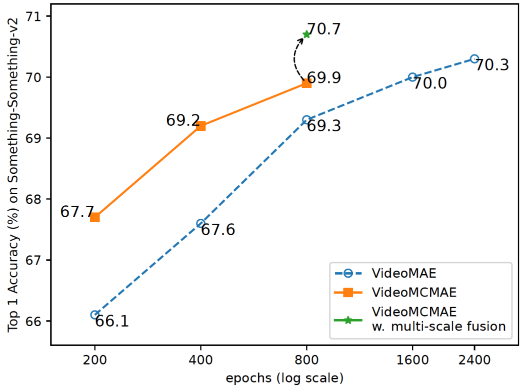
Methods	Pretraining	P-Epochs	F-Epochs	AP^{box}	AP^{mask}	Params (M)	FLOPs (T)
Benmarking [37]	IN1K w/o labels	1600	100	50.3	44.9	118	0.9
ViTDet [35]	IN1K w/o labels	1600	100	51.2	45.5	111	0.8
MIMDET [20]	IN1K w/o labels	1600	36	51.5	46.0	127	1.1
Swin+ [42]	IN1K w/ labels	300	36	49.2	43.5	107	0.7
MViTv2 [36]	IN1K w/ labels	300	36	51.0	45.7	71	0.6
MCMAE	IN1K w/o labels	1600	25	53.2	47.1	104	0.9

Semantic Segmentation with UperNet (ADE20k)

Models	Pretrain Data	P-Epochs	mIoU	Params (M)	FLOPs (T)
DeiT-B [51]	IN1K w/ labels	300	45.6	163	0.6
Swin-B [42]	IN1K w/ labels	300	48.1	121	0.3
MoCo V3 [29]	IN1K	300	47.3	163	0.6
DINO [6]	IN1K	400	47.2	163	0.6
BEiT [2]	IN1K+DALLE	1600	47.1	163	0.6
PeCo [17]	IN1K	300	46.7	163	0.6
CAE [9]	IN1K+DALLE	800	48.8	163	0.6
MAE [28]	IN1K	1600	48.1	163	0.6
MCMAE	IN1K	1600	51.7	153	0.6

Video Recognition (K400 & SSv2)

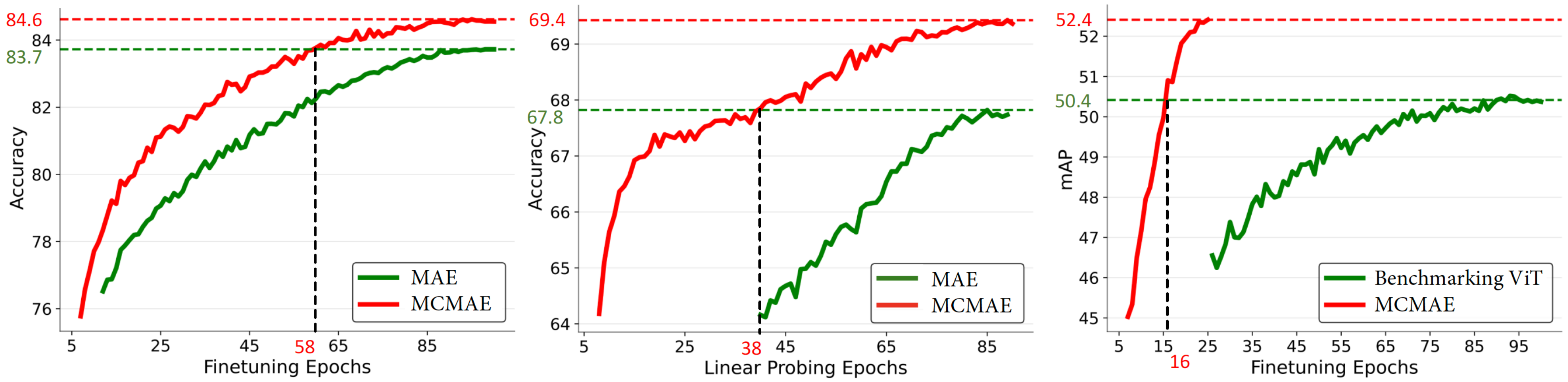




NEURAL INFORMATION PROCESSING SYSTEMS

Experiments – Faster Convergence

- MCMAE converges faster on downstream tasks



(a) ImageNet Finetuning (b) ImageNet Linear Probing (c) COCO Detection

Experiments – Ablation Studies

NEURAL INFORMATION PROCESSING SYSTEMS

Ablation studies of masked conv, block-wise masking, and conv kernel size

P-Epochs	Masked Conv	Block Masking	5×5 Conv	7×7 Conv	9×9 Conv	FT (%)	FLOPs
800	✓	✓	✓	✗	✗	84.6	1×
	✓	✗	✓	✗	✗	84.2	1.7×
	✗	✓	✓	✗	✗	81.5	1×
	✓	✓	✓	✗	✗	84.5	0.997×
	✓	✓	✗	✓	✗	84.4	1.003×
	✓	✓	✗	✗	✓	84.6	1.007×

Ablation studies of pretraining epochs

Pretrain Epochs	ImageNet		COCO		ADE20K
	FT	LIN	AP^{box}	AP^{mask}	mIoU
200	84.1	62.5	50.2	44.8	48.1
400	84.4	66.9	51.4	45.7	49.5
800	84.6	68.4	52.0	46.3	50.2
1600	84.6	69.4	52.5	46.5	50.7

Ablation studies of multi-scale decoder

P-Epochs	Method	FT (%)	LIN (%)	AP^{box}	AP^{mask}	mIoU
200	MCMAE-Base	84.1	N/A	50.2	44.8	48.1
	w/ multi-scale decoder	84.4	N/A	50.8	45.4	48.5
1600	MCMAE-Base	84.6	69.4	52.5	46.5	50.7
	w/ multi-scale decoder	85.0	70.9	53.2	47.1	51.7

Thanks!

For further questions welcome to discuss via GitHub issues

https://github.com/Alpha-VL/ConvMAE

NEURAL INFORMATION PROCESSING SYSTEMS