ConvMAE: Masked Convolution Meets Masked Autoencoders

Peng Gao[1] **Teli Ma**[1] **Hongsheng Li**[1,2] **Ziyi Lin**[2] **Jifeng Dai**[3] **Yu Qiao**[1]
[1] Shanghai AI Laboratory [2] MMLab, CUHK
[3] SenseTime Research

NeurIPS 2022

分享人：高逸凡

目录

作者介绍

Yu Qiao

FOLLOW

Professor of Shanghai AI Laboratory; Shenzhen Institutes of Advanced Technology, CAS
Verified email at siat.ac.cn - Homepage
Computer Vision   Deep Learning   Speech Processing   Pattern Recognition

TITLE   CITED BY   YEAR

GET MY OWN PROFILE

Cited by   VIEW ALL

	All	Since 2018
Citations	42047	35596
h-index	79	72

Jifeng Dai

关注

Associate Professor of EE, Tsinghua University; Adjuct Researcher of Shanghai AI Laboratory
在 tsinghua.edu.cn 的电子邮件经过验证 - 首页
computer vision   deep learning

标题   引用次数   年份

创建我的个人资料

引用次数   查看全部

	总计	2018 年至今
引用	25385	23880
h 指数	33	33

Hongsheng Li (李鸿升)

关注

Associate Professor at The Chinese University of Hong Kong
在 ee.cuhk.edu.hk 的电子邮件经过验证 - 首页
Computer Vision   Machine Learning   Medical Image Analysis

标题   引用次数   年份

创建我的个人资料

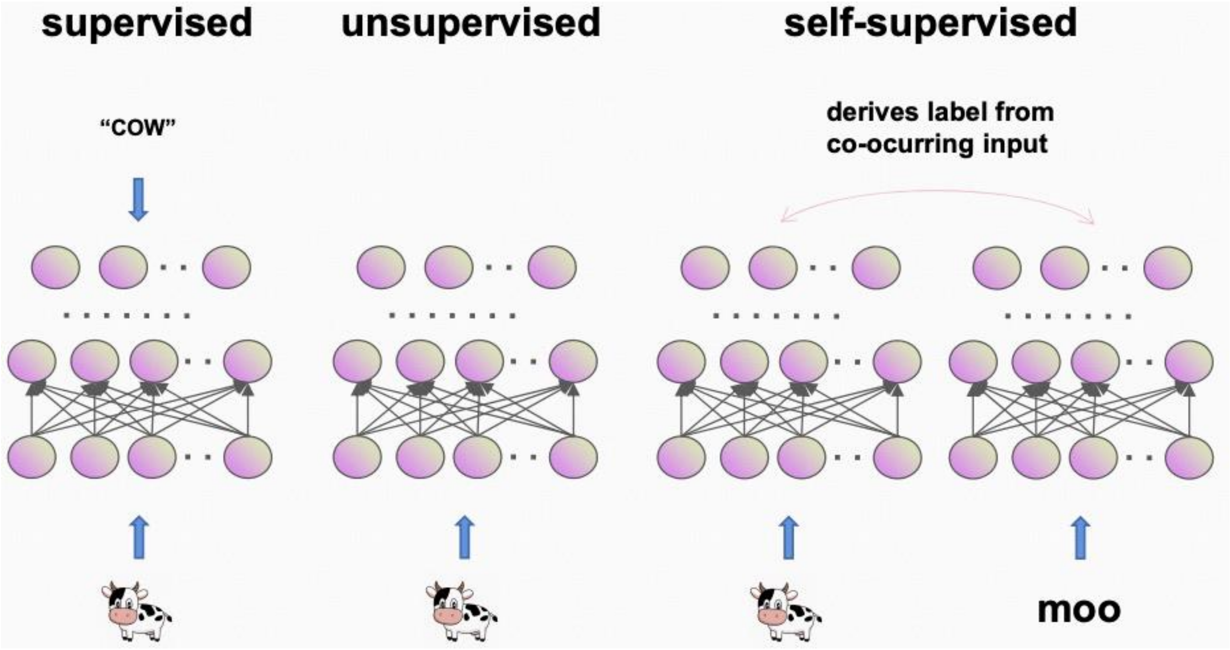
引用次数   查看全部

	总计	2018 年至今
引用	26099	23541
h 指数	74	70

- 作者介绍
- 研究背景
- 研究动机
- 本文方法
- 实验效果
- 总结反思

研究背景

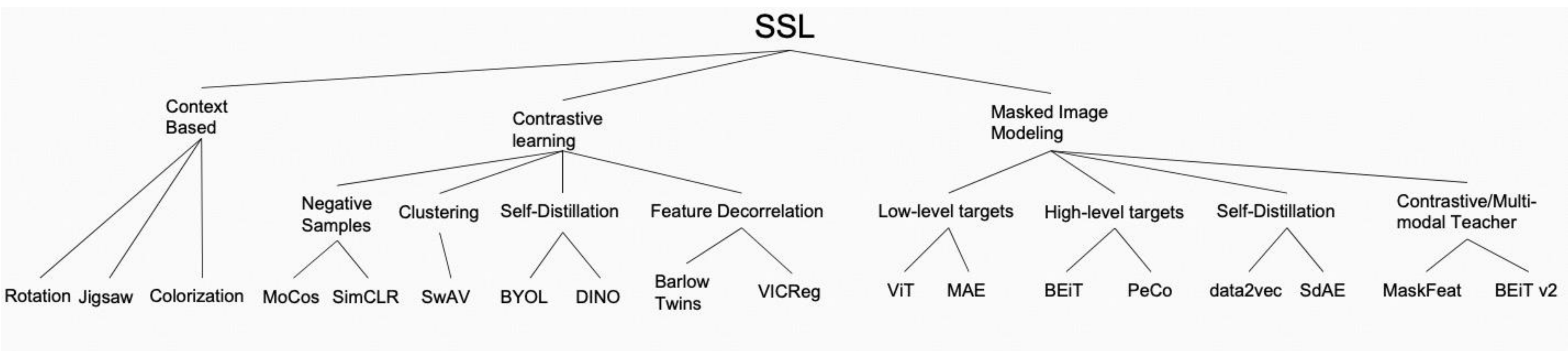
- 自监督学习
  - 一种基于pretext task的无监督学习范式



[1] de Sa V R. Learning classification with unlabeled data[J]. Advances in neural information processing systems, 1994: 112-112.

研究背景

- 自监督学习
  - 依据pretext task划分自监督学习的种类
    - 基于上下文
    - 基于对比学习
    - 基于掩码图像建模（生成式）

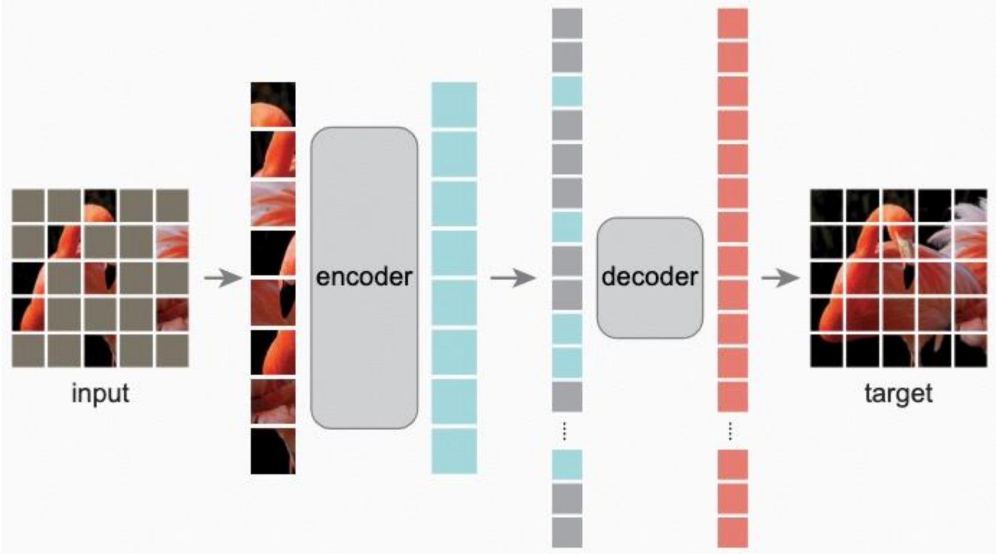


[2] Gui J, Chen T, Cao Q, et al. A Survey of Self-Supervised Learning from Multiple Perspectives: Algorithms, Theory, Applications and Future Trends[J]. arXiv preprint arXiv:2301.05712, 2023.

- 作者介绍
- 研究背景
- 研究动机
- 本文方法
- 实验效果
- 总结反思

研究动机

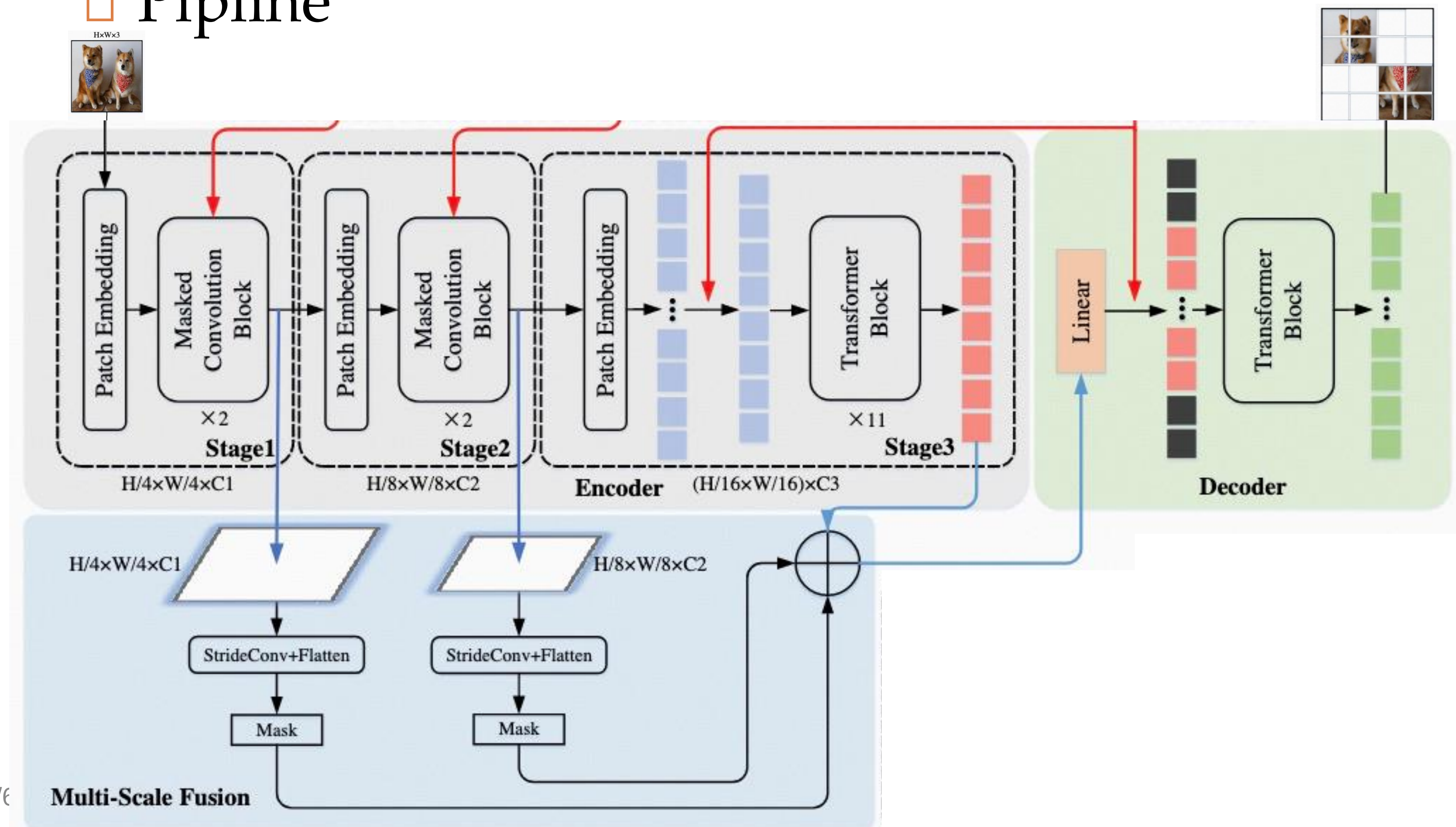
- 生成式自监督学习
  - MAE（Masked Autoencoders ）
    - 存在pretraining-finetuning 差异
    - 没有multi-scale特征
    - 能否用CNN?



- 作者介绍
- 研究动机
- 本文方法
- 实验效果
- 总结反思

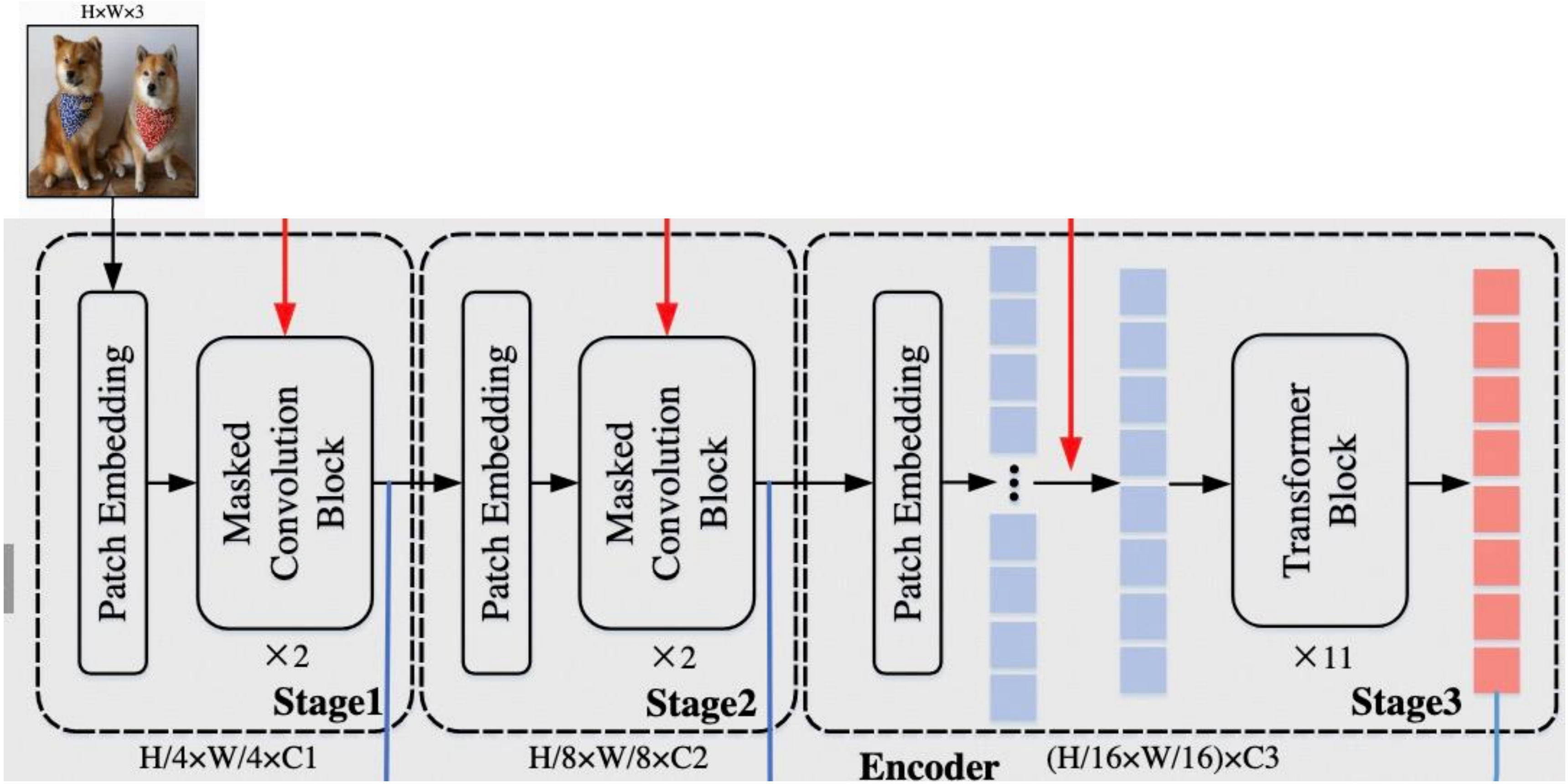
本文方法

- Pipline



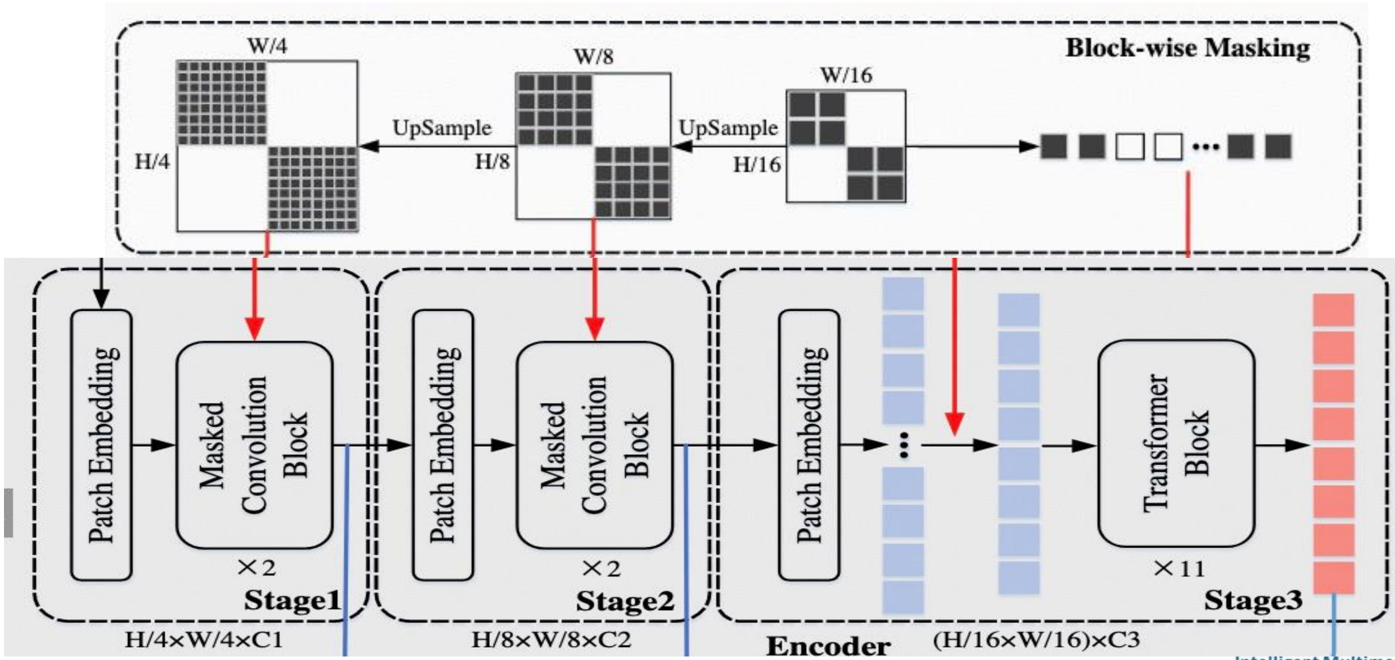
本文方法

- The Hybrid Convolution-transformer Encoder



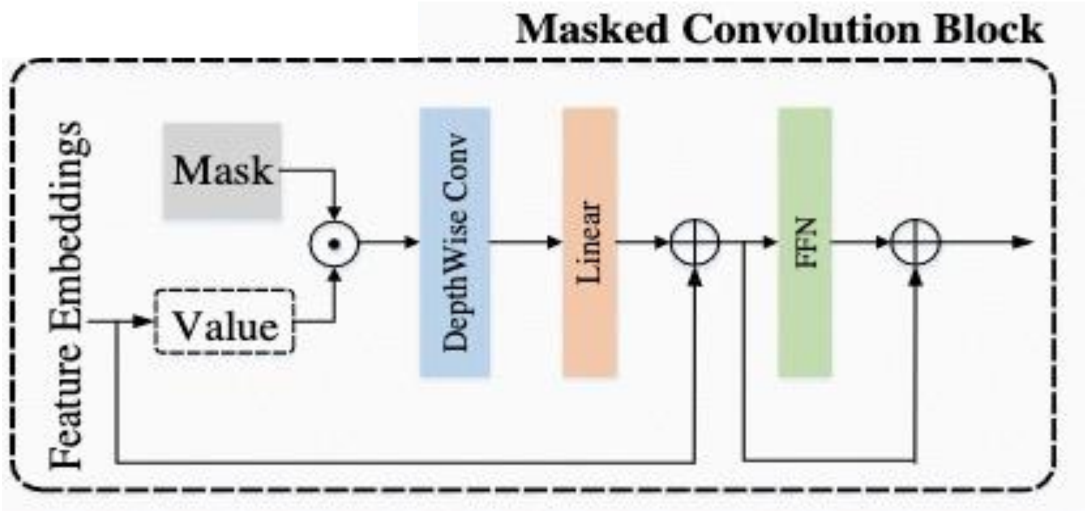
本文方法

- Block-wise Masking with Masked Convolutions
  - Block-wise Masking
    - 先产生stage 3的 mask (random mask ratios 75%)
    - 再UpSample到stage 2、 stage 2



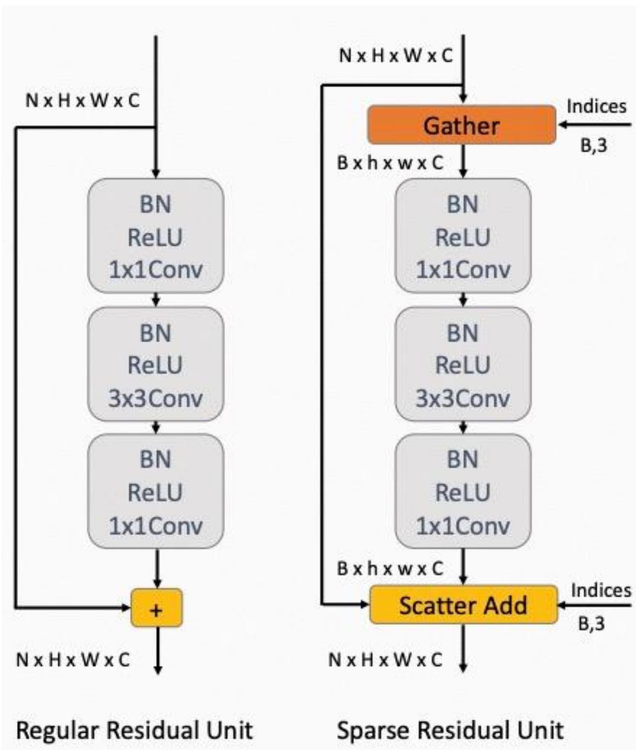
本文方法

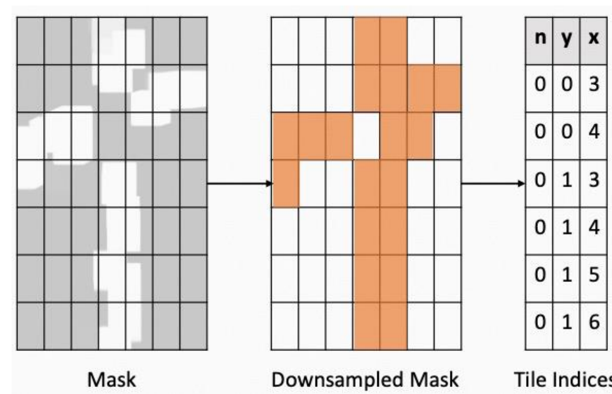
- Block-wise Masking with Masked Convolutions
  - Masked Convolutions Block
    - 采用mask卷积，避免信息泄露

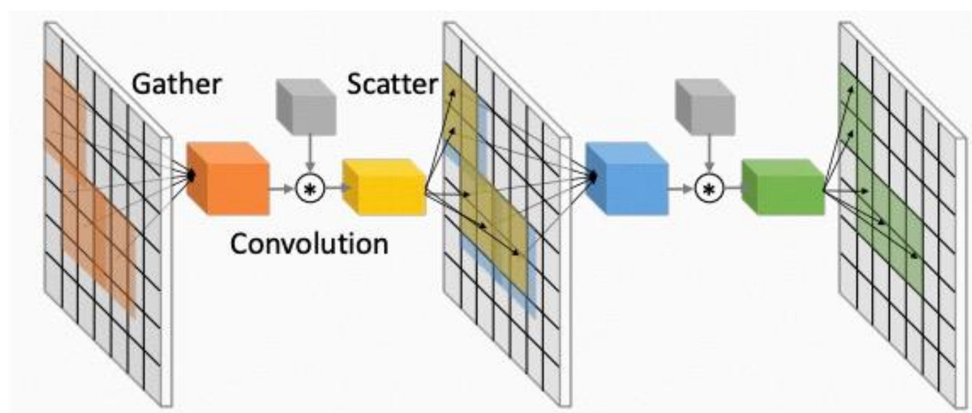


本文方法

- Block-wise Masking with Masked Convolutions
  - Masked Conv（tiled sparse convolution）

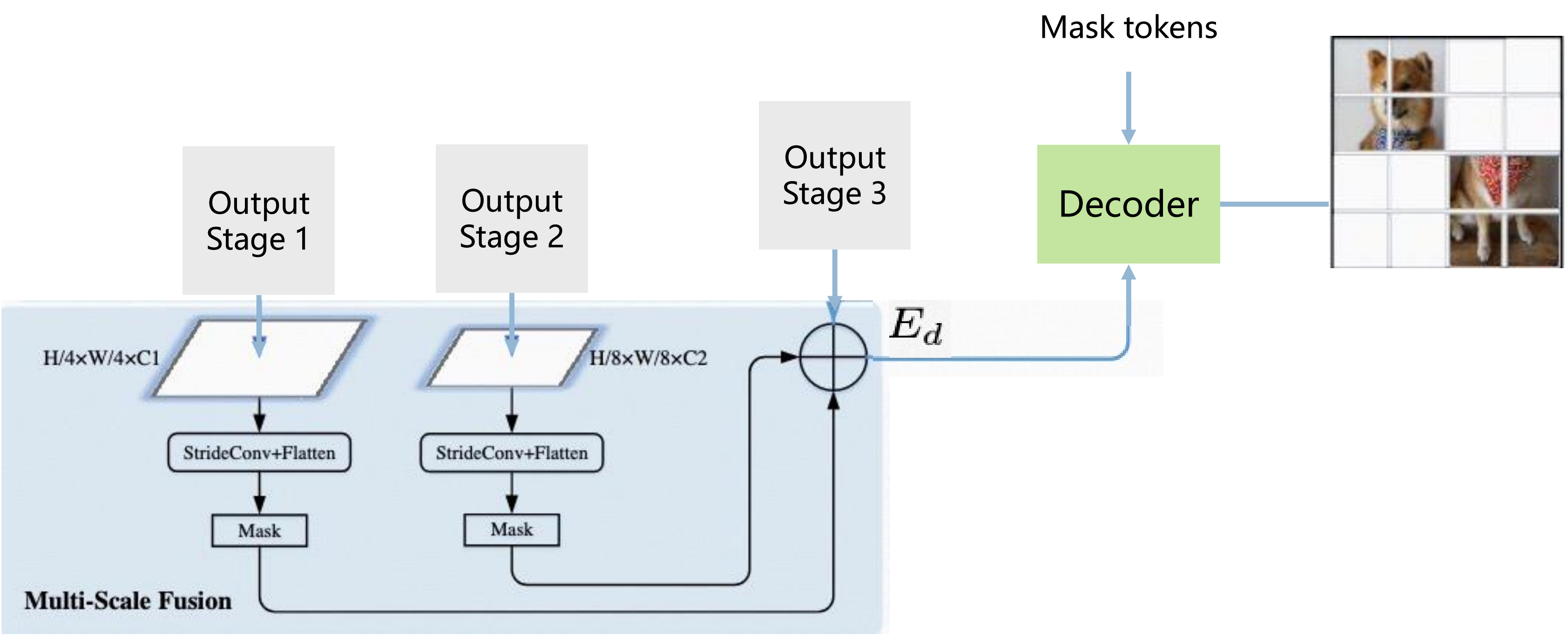






本文方法

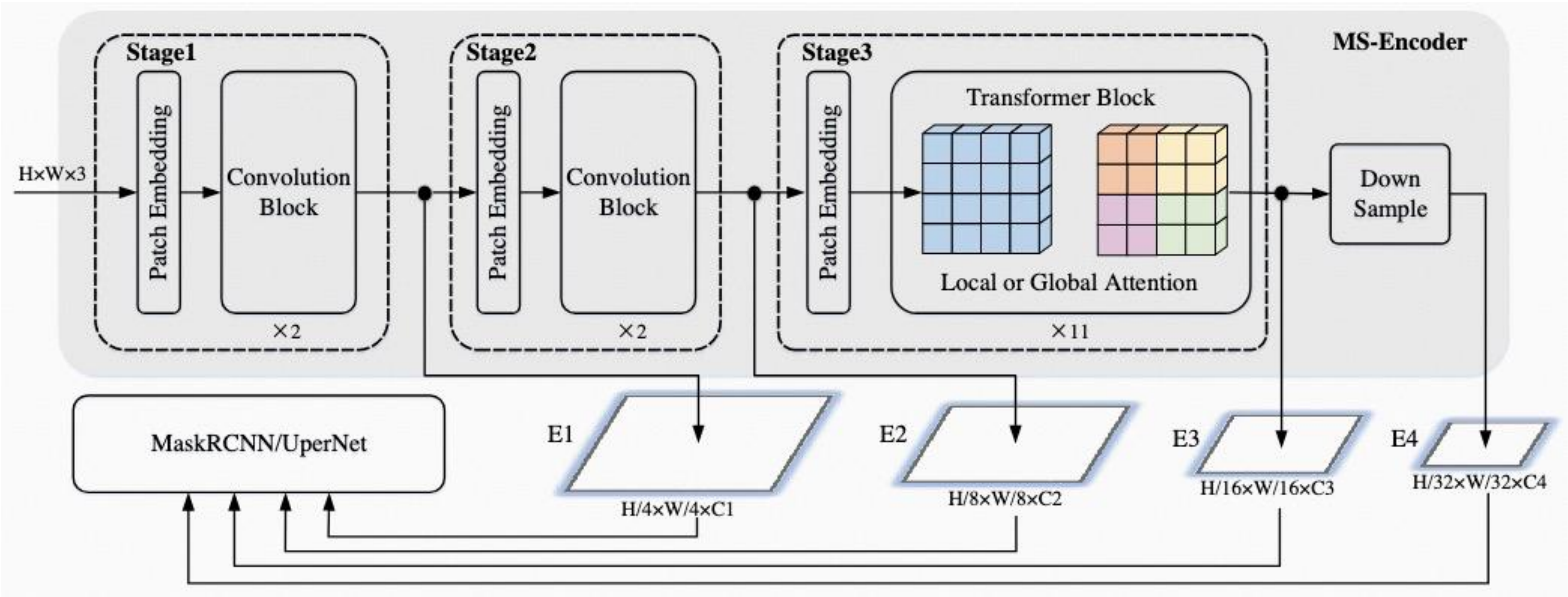
- The Multi-scale Decoder and Loss



$$E_d = \mathrm{Linear}(\mathrm{StrideConv}(E_1, 4) + \mathrm{StrideConv}(E_2, 2) + E_3),$$

本文方法

- ConMAE用于目标检测和图像分割



- 作者介绍
- 研究动机
- 本文方法
- 实验效果
- 总结反思

实验效果

- ImageNet-1K Finetune和Linear probe

Results on ImageNet-1K Finetuning. We report the accuracy of ConvMAE on Table 1 and conduct

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
ConvMAE	ConViT-B	88	RGB	25%	1600	85.0	70.9

Table 1: Comparison with state-of-the art mask auto-encoding schemes with similar model size. FT and LIN denotes ImageNet-1K finetuning and linear probe accuracy respectively.

实验效果

- Object Detection and Semantic Segmentation

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
ConvMAE	IN1K w/o labels	1600	25	53.2	47.1	104	0.9

Table 2: Performances of different pretrained backbones on object detection with Mask-RCNN [30].

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
ConvMAE	IN1K	1600	51.7	153	0.6

Table 3: Comparison with different pretrained backbones on ADE20k with UperNet.

实验效果

- Video Understanding

Pretrain Epochs	ImageNet FT	ImageNet LIN	COCO AP^{box}	COCO AP^{mask}	ADE20K 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

Table 4: The influence of increasing pretraining epochs on various downstream tasks.

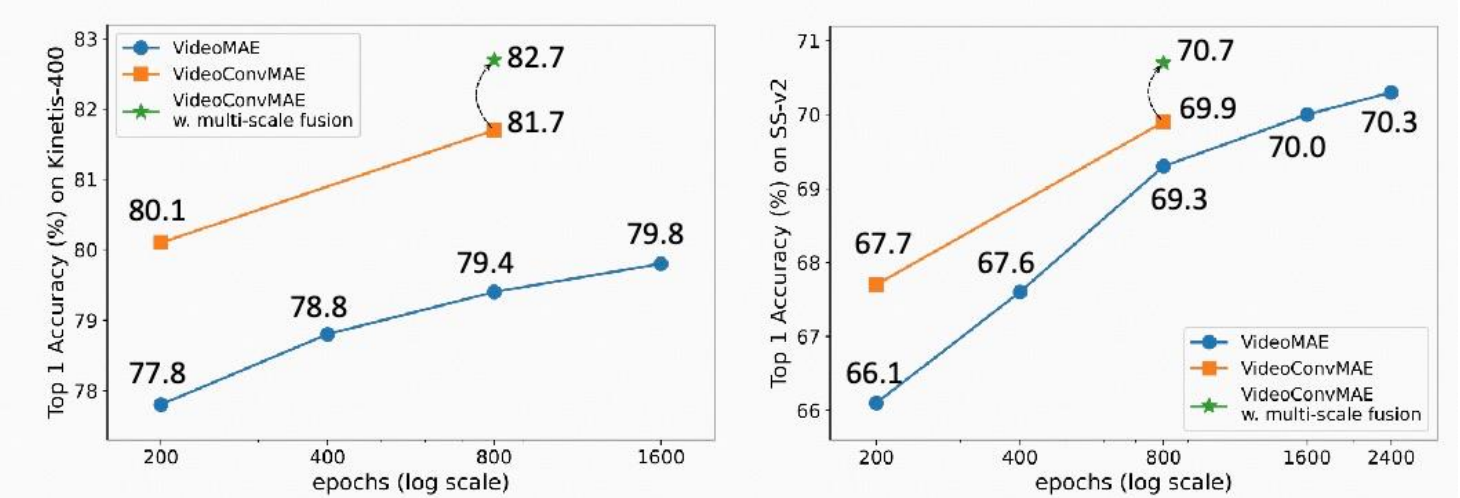


Figure 3: Finetuning accuracy on Kinetics-400 and Something-Something-v2.

实验效果

- Ablation Study
  - 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

Table 4: The influence of increasing pretraining epochs on various downstream tasks.

实验效果

- Ablation Study
  - Input-token random maskgin
  - Influence of masked convolution
  - kernel sizes in stages 1 and 2

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×

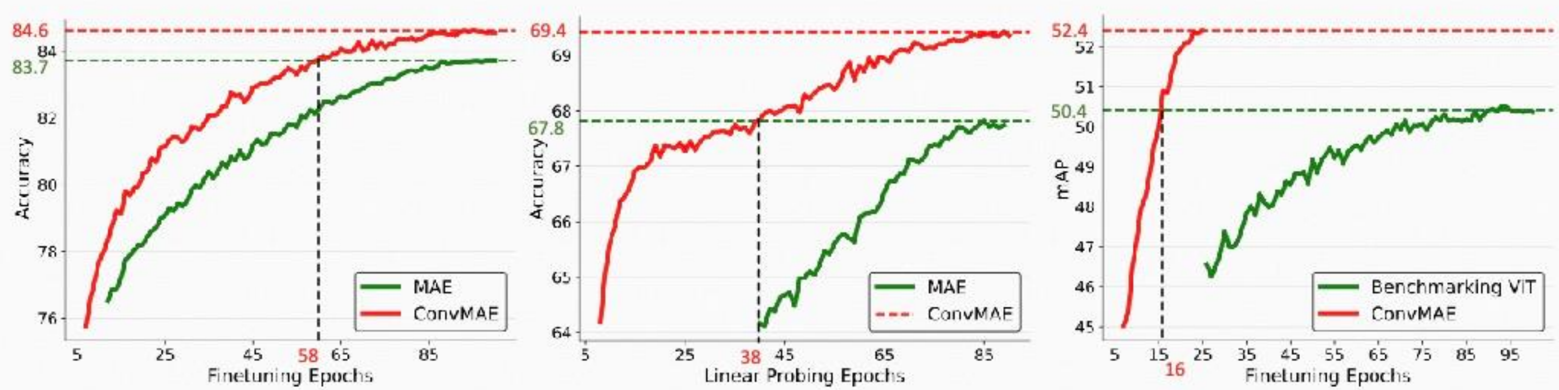
Table 5: Ablation study on the influence of the masked conv, block masking, kernel size in stages 1 and 2 of ConvMAE on ImageNet-1K finetuning accuracy.

实验效果

- Ablation Study
  - Multi-scale Decoder
  - Convergence speed

P-Epochs	Method	FT (%)	LIN (%)	AP^{box}	AP^{mask}	mIoU
200	ConvMAE-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	ConvMAE-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

Table 6: For a base ConvMAE pretrained for 200 epochs and 1600 epochs, we ablate the multi-scale decoder on ImageNet finetuning and object detection on COCO.



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

总结反思

- 效果提升明显，Pipline略复杂
- 在自监督的基础上学习多尺度特征效果明显

Thank for your attention !