视觉自监督学习年度进展评述

胡瀚
微软亚洲研究院
@VALSE 2023 无锡
2023.06.10

视觉预训练

- “预训练-微调”——计算机视觉领域的重要范式

预训练

迁移
预训练模型

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

CITYSCAPES
DATASET

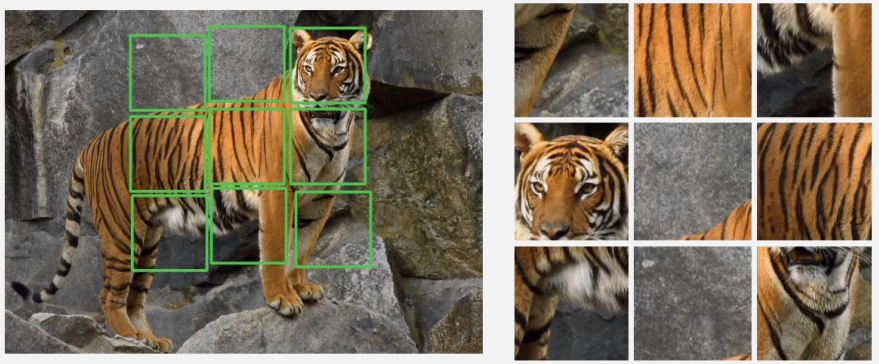
COCO
Common Objects in Context

在下游任务微
- 细粒度分类
- 物体检测
- 语义分割

- 从有监督到**自监督**

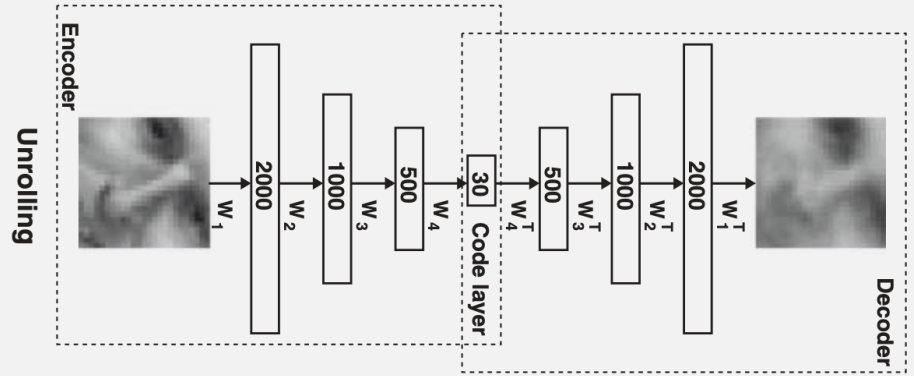
自监督预训练

- 监督信号来自数据本身，从数据出发构造预训练/前置任务



拼图游戏

旋转预测



自编码器

为什么需要自监督预训练?

蛋糕类比

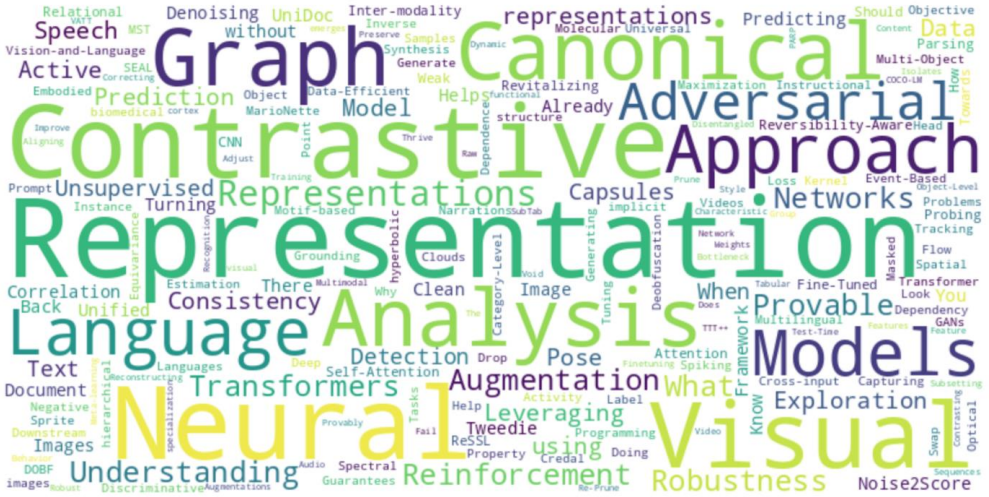
- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - A few bits for some samples
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - 10→10,000 bits per sample
- 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

图片来源：Yann LeCun

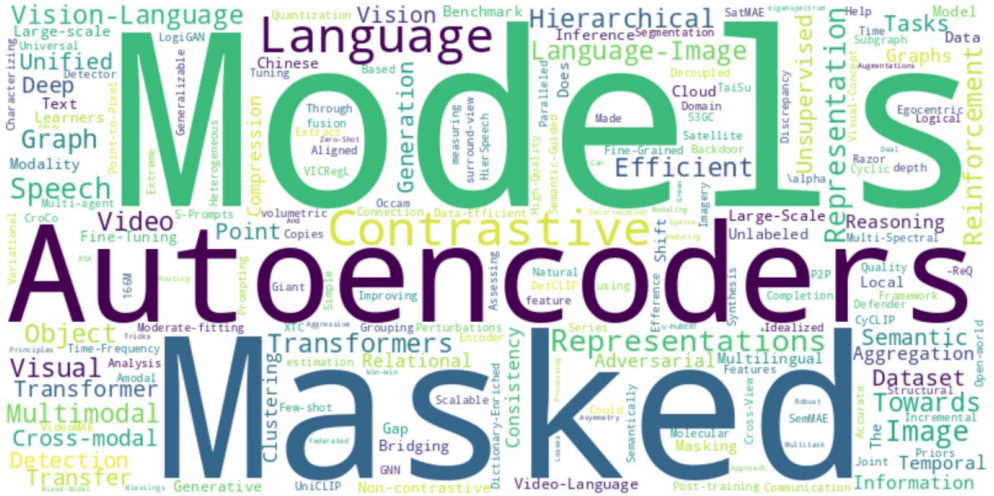
婴儿期学习方式

图片来源：Andrew Zisserman

2022年视觉自监督学习的主旋律：Masked



NeurIPS 2021



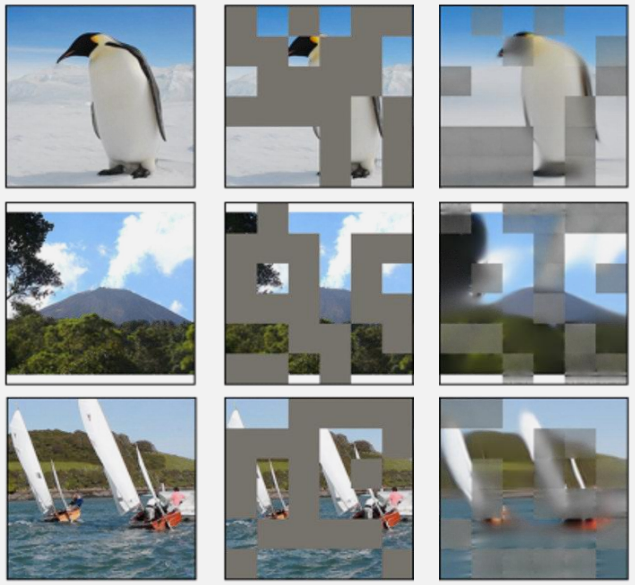
NeurIPS 2022

CVPR 2022

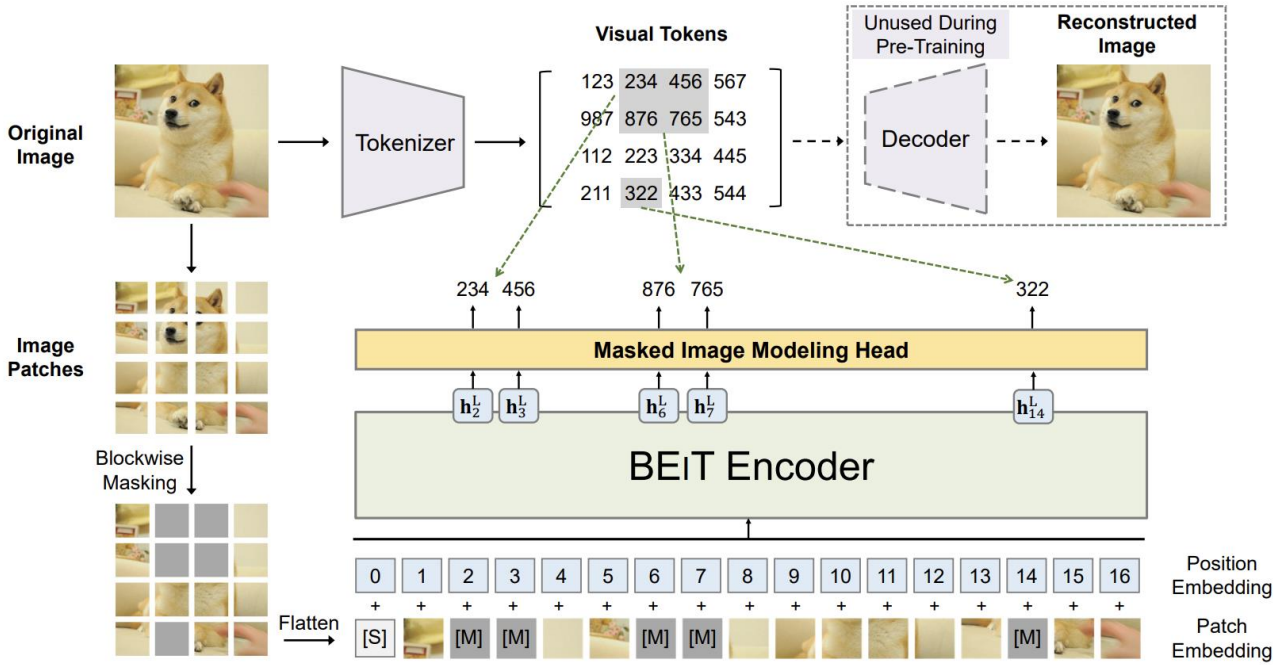
CVPR 2023

掩码图像建模（Masked Image Modeling）

- BEiT [微软, ICLR2021]
  - 掩盖住一部分图像像素，训练神经网络去恢复被掩住的内容
  - **微调性能取得突破**



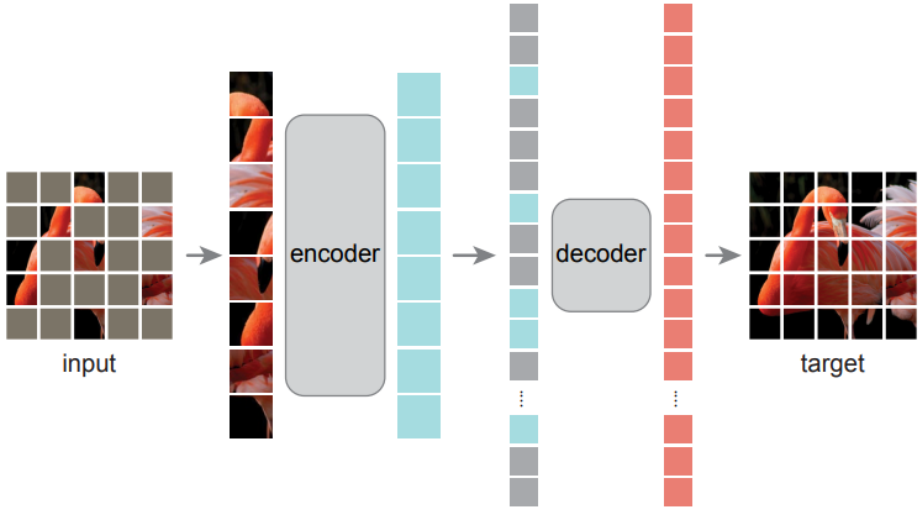
预训练任务: 掩码图像建模(MIM)



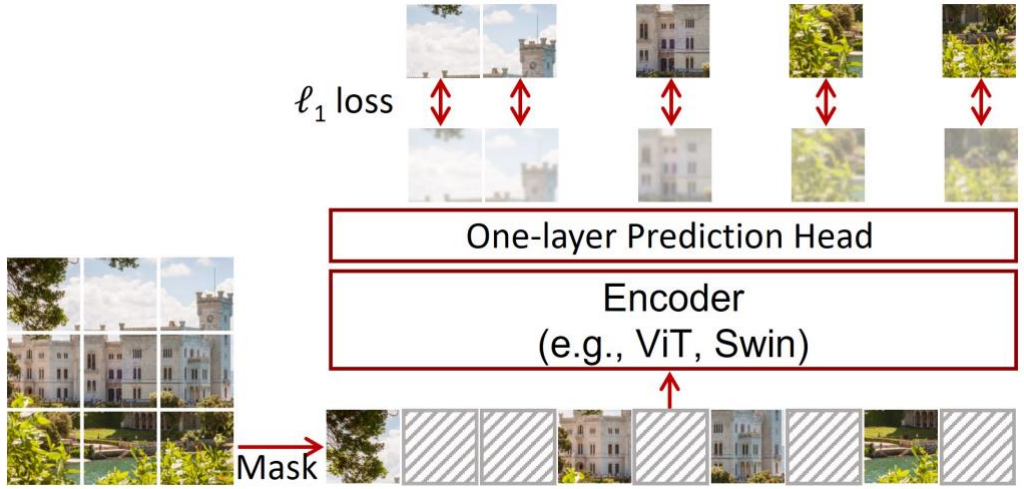
BEiT [微软, ICLR2021]

掩码图像建模（MIM）引起主流关注

- iBOT [字节2021.10]
- MAE[Meta2021.11]
- SimMIM [微软2021.11]
- MaskFeat [Meta2021.12]
- PeCo[微软2021.12]



MAE



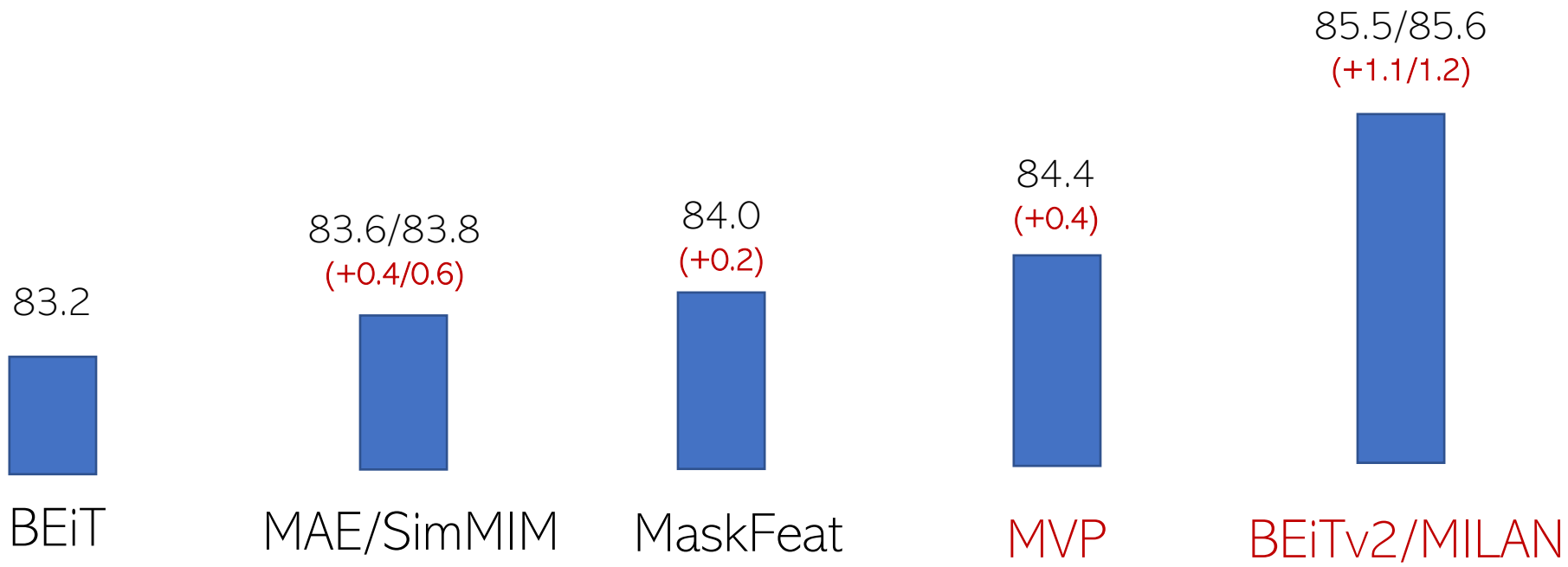
SimMIM

自监督学习年度进展（2022-2023）

- 技术进展趋势一：掩码图像建模的**改进**
- 技术进展趋势二：发现掩码图像建模对**大模型**比较友好
- 技术进展趋势三：针对**小模型**的掩码图像建模训练
- 技术进展趋势四：挖掘掩码图像建模的**好性质**
- 技术进展趋势五：**拓展到其它模态**

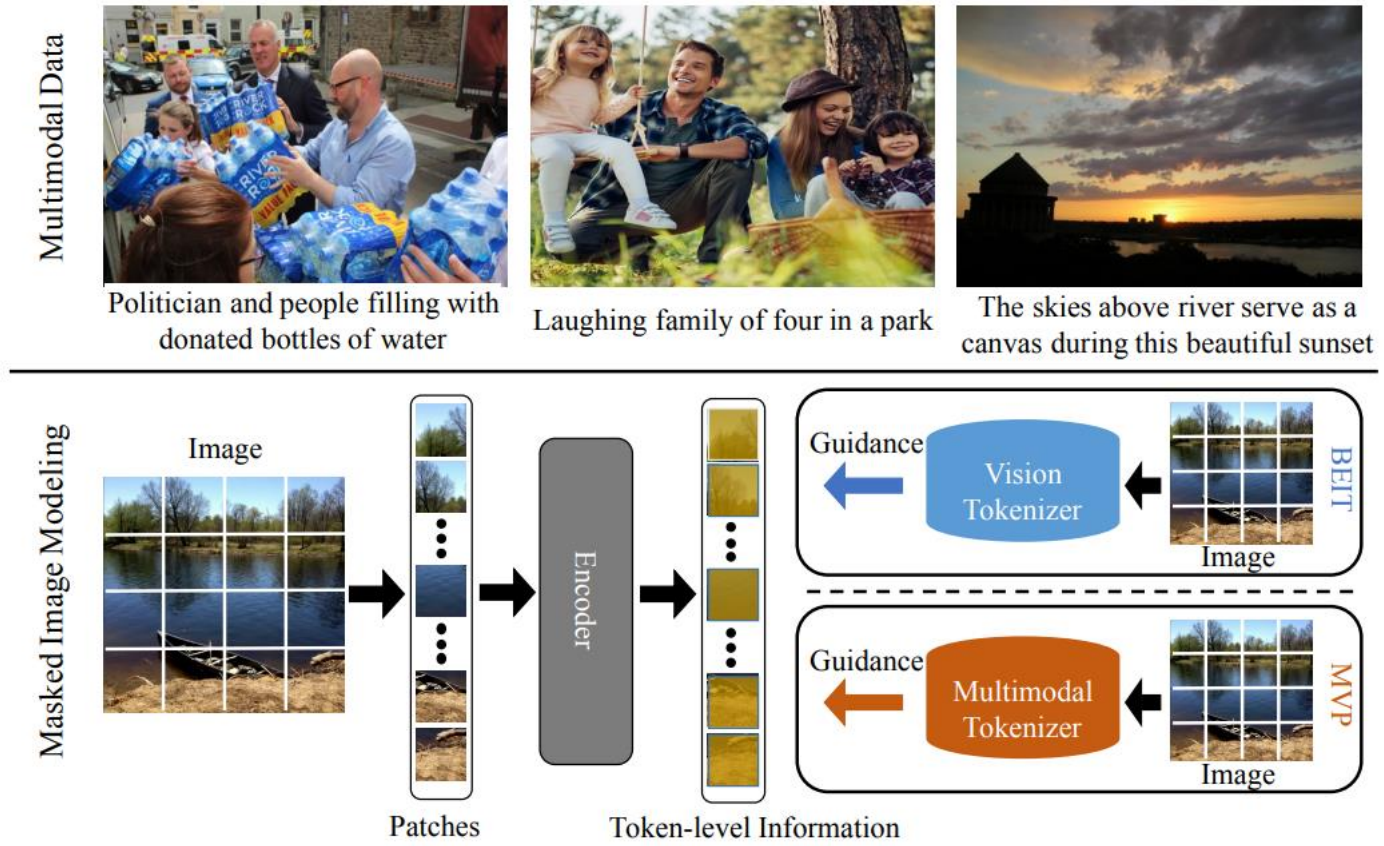
技术进展一：掩码图像建模的改进

- 性能改进（以ViT-B骨干网络，IN-1K top-1准确率为例）
  - 83.2 -> 85.6 **(+2.4)**



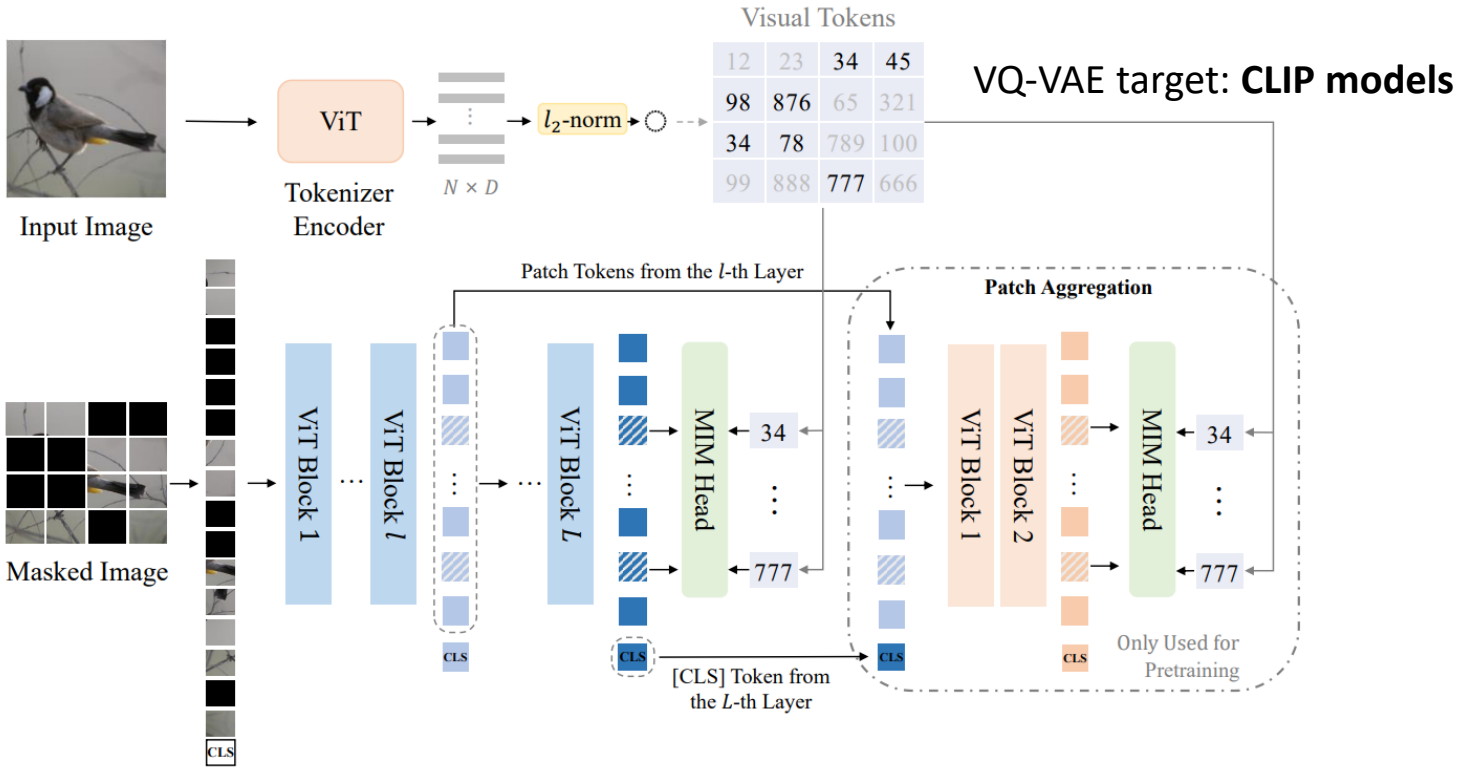
技术进展一：掩码图像建模的改进

- 以CLIP模型特征为重构目标
  - MVP[中科大&华为]



技术进展一：掩码图像建模的改进

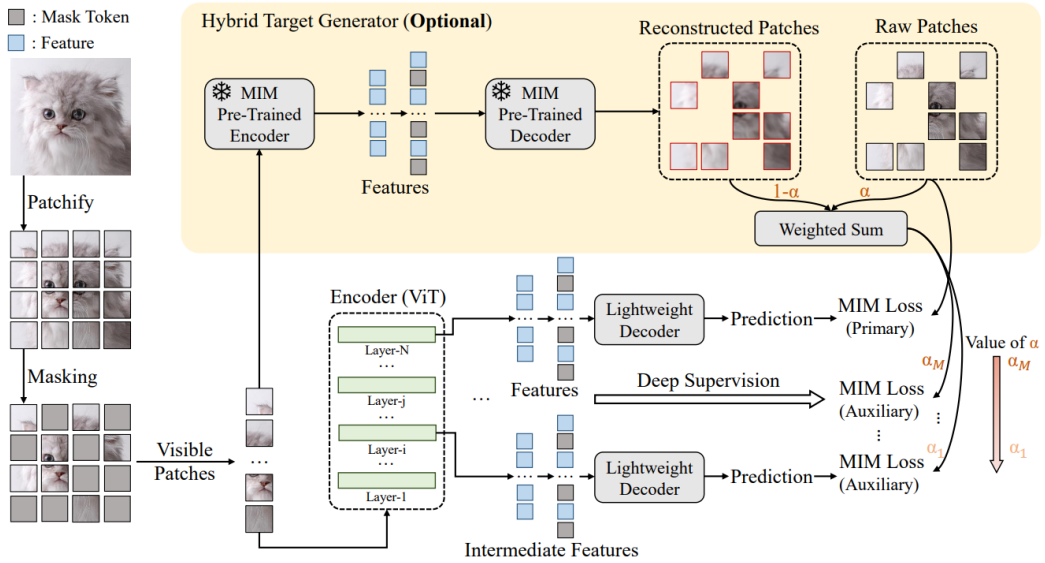
- 以CLIP模型特征为重构目标
  - BEiT-V2 [微软]
  - MILAN [普林斯顿&阿里]

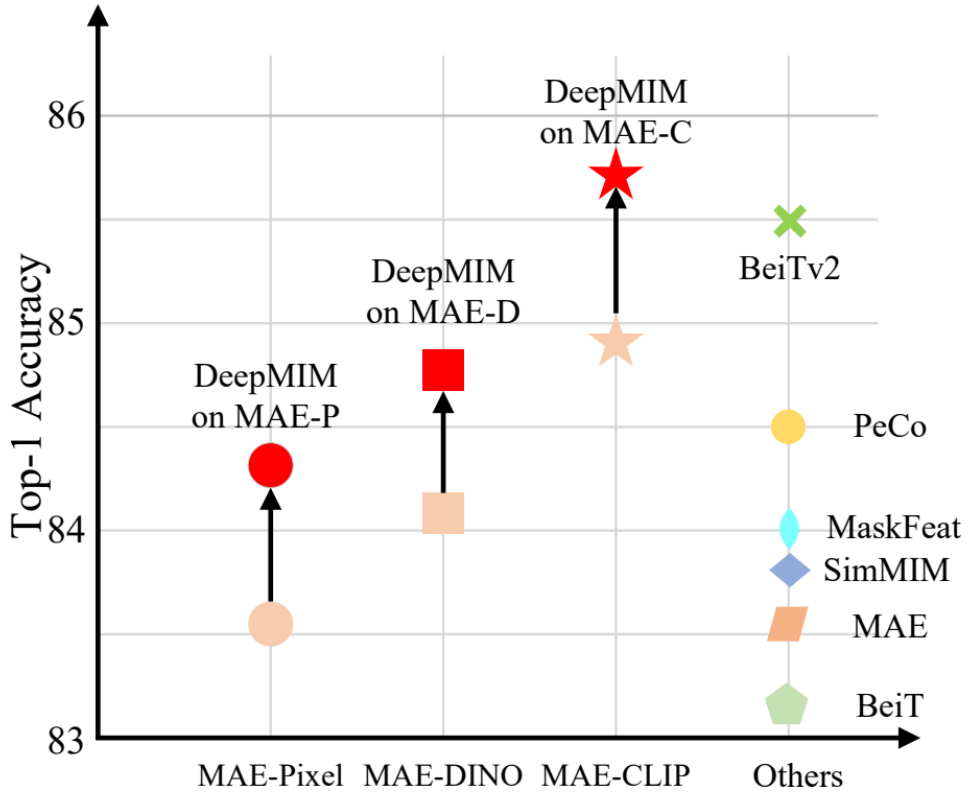


BEiT-V2

技术进展一：掩码图像建模的改进

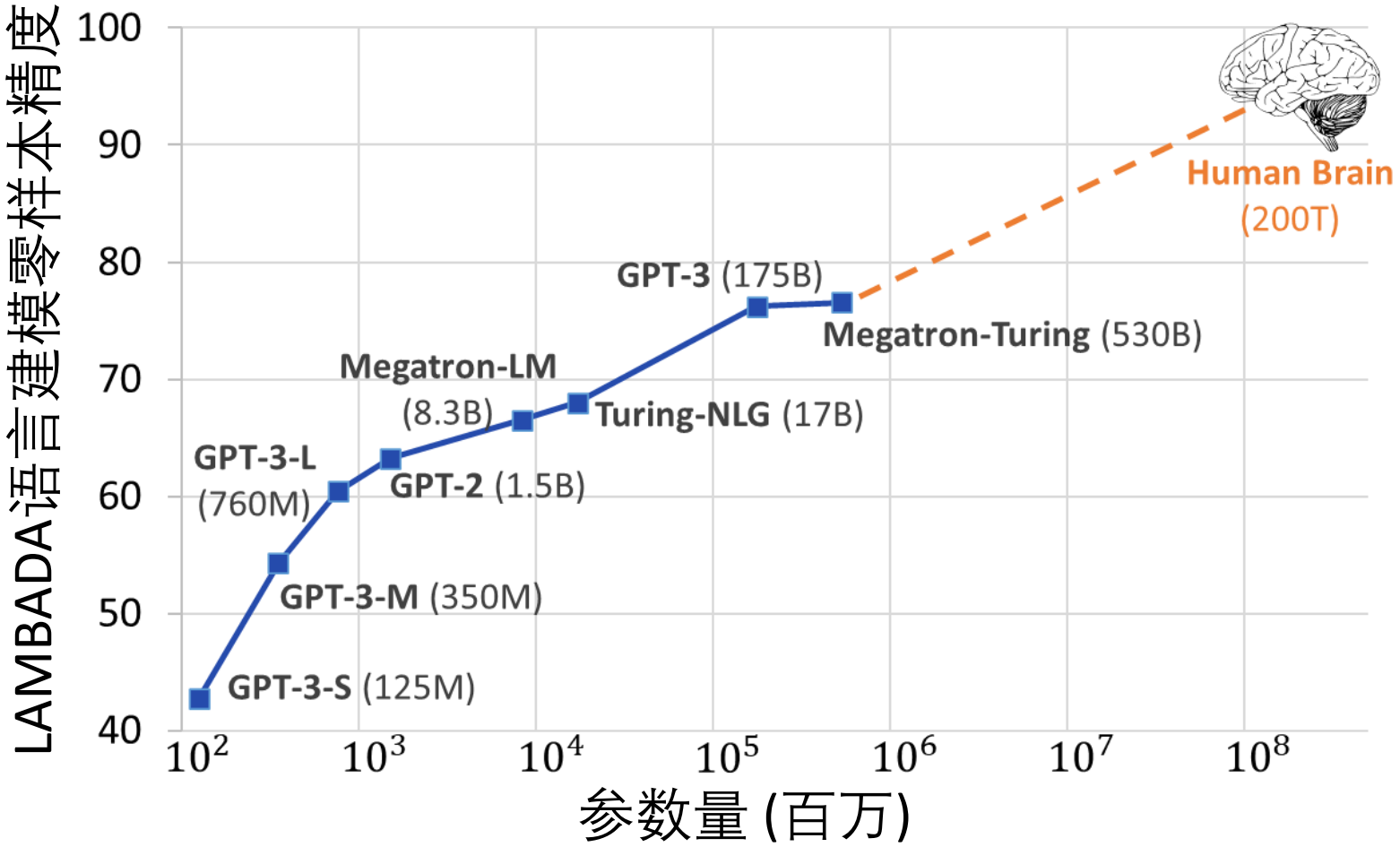
- 其它改进
  - 深度监督: DeepMIM [微软]





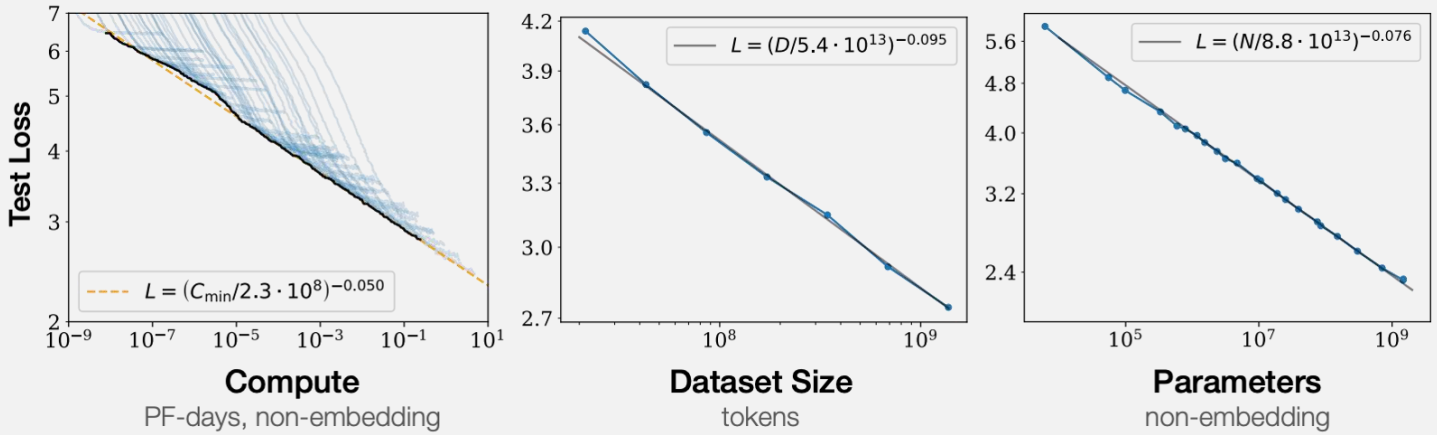
技术进展二：发现掩码图像建模对大模型比较友好

- NLP模型容量的扩展能持续改进NLP任务的性能

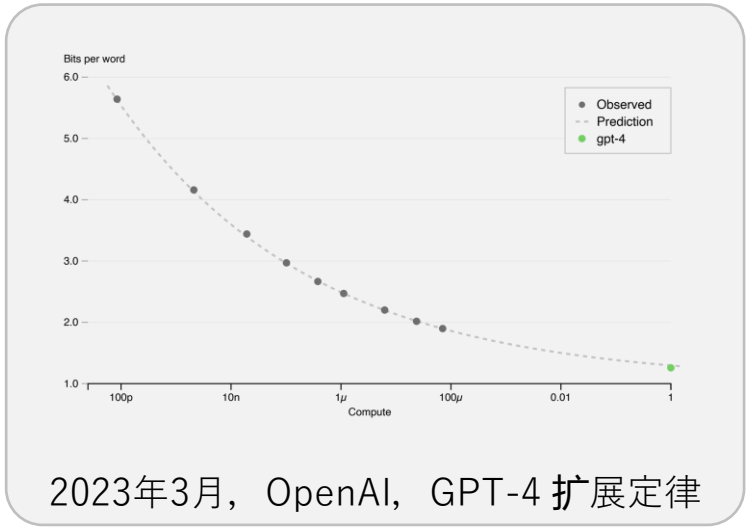


技术进展二：发现掩码图像建模对大模型比较友好

- 指引NLP大模型的扩展定率（scaling law）



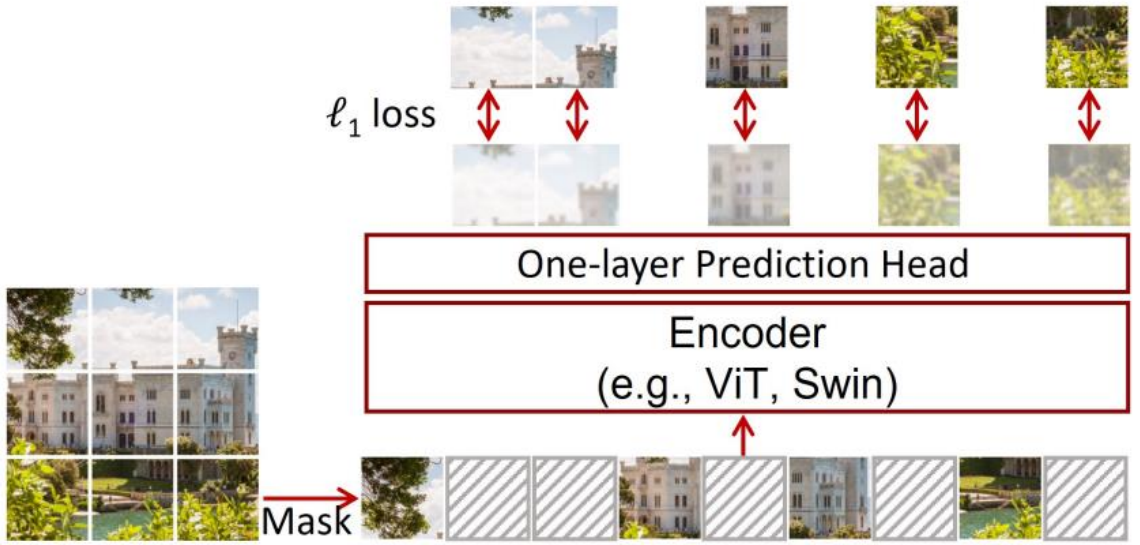
2020年1月，OpenAI，Scaling Laws for Neural Language Models



2023年3月，OpenAI，GPT-4 扩展定律

技术进展二：发现掩码图像建模对大模型比较友好

- Swin V2[中科大&微软, CVPR2022]：首个基于掩码图像建模的10亿级参数视觉大模型（30亿参数）
  - 比基于图像分类任务的ViT-G模型（18亿参数）训练代价**小10倍**，需要的标注数据**小40倍**



Swin V2-G基于掩码图像建模方法预训练得到

技术进展二：发现掩码图像建模对大模型比较友好

- CAE v2 [百度]
- EVA [智源]
- InternImage [商汤]

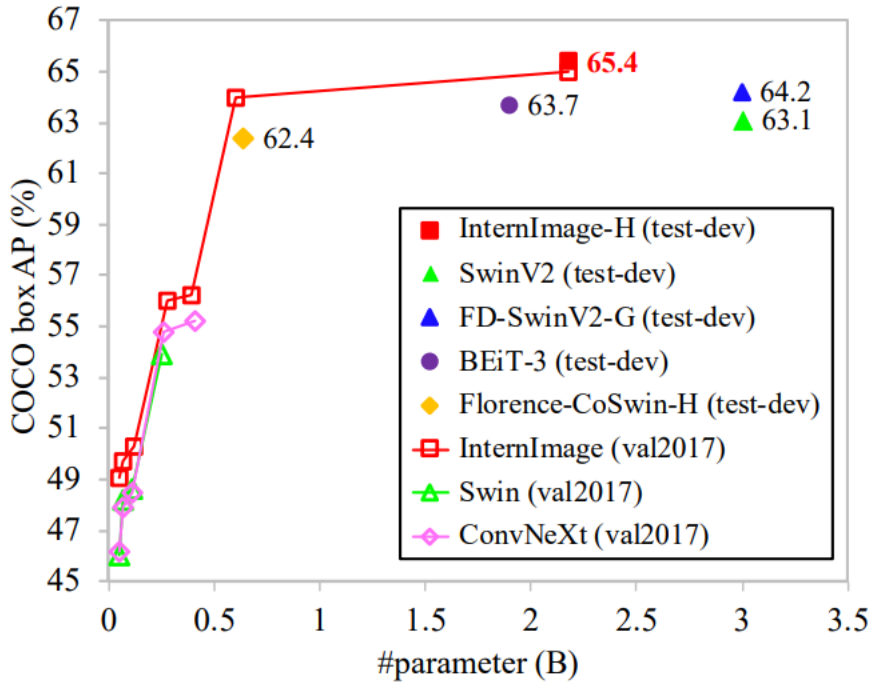


Figure 2. **Performance comparison on COCO of different backbones.** The proposed InternImage-H achieves a new record 65.4 box AP on COCO test-dev, significantly outperforming state-of-the-art CNNs and large-scale ViTs.

InternImage

技术进展二：发现掩码图像建模对大模型比较友好

- 为什么?
  - 信息量大，能挖掘大模型的潜力

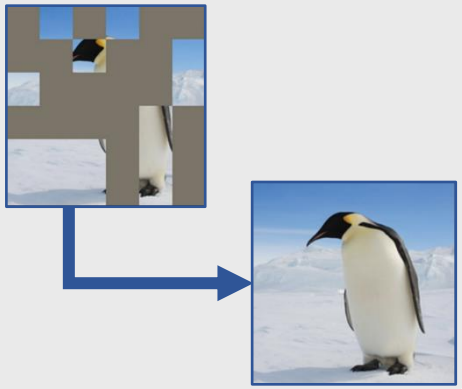
分类

1,00000类
≈ 17 Bits

视觉-语言对比学习
(E.g., CLIP)

1,000,000个句子
≈ 20 Bits

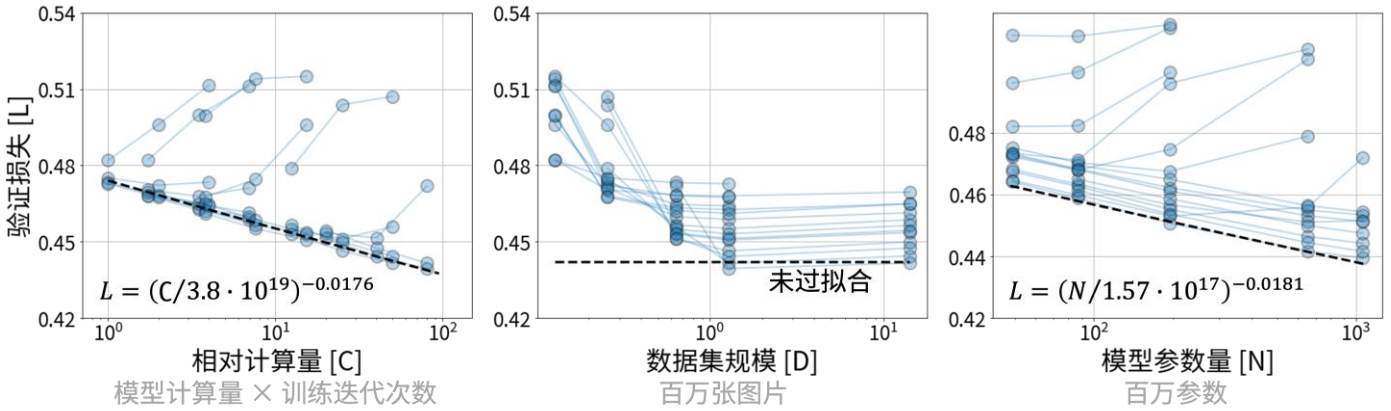
掩码图像建模



224x224像素
>> **100,000 Bits**

技术进展二：发现掩码图像建模对大模型比较友好

- 在大数据上的表现
  - Data Scaling [清华&微软 CVPR22]
  - MAE on Billion-data [Meta Arxiv2023]



Data Scaling [清华&微软 CVPR22]

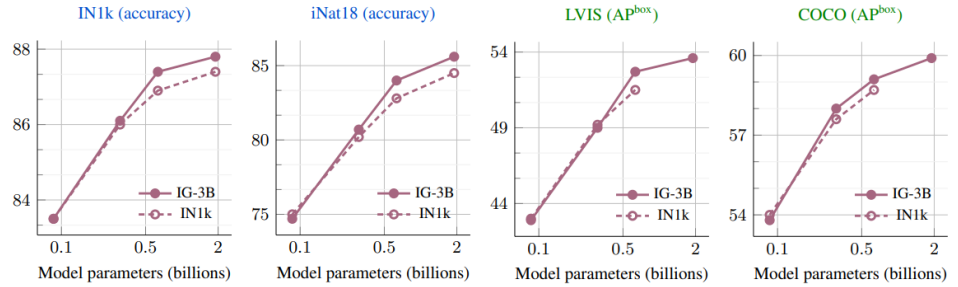


Figure 2: Scaling MAE with model and dataset size. We plot MAE's performance when pretrained on ImageNet-1k or Instagram-3B and finetuned on downstream tasks. MAE scales to billion parameters sized models using just IN1k pretraining. Larger models show improved scaling behavior when pretrained with the much larger IG-3B dataset. MAE pretrained on IN1k data point is missing for the 2 billion model as training at that scale was unstable on both COCO and LVIS datasets.

MAE on Billion-data [Meta Arxiv2023]

技术进展三：针对小模型的掩码图像建模训练

- 主要思想：蒸馏
  - TinyMIM [微软, CVPR2023]
  - Lightweight MAE [自动化所，Arxiv2023]

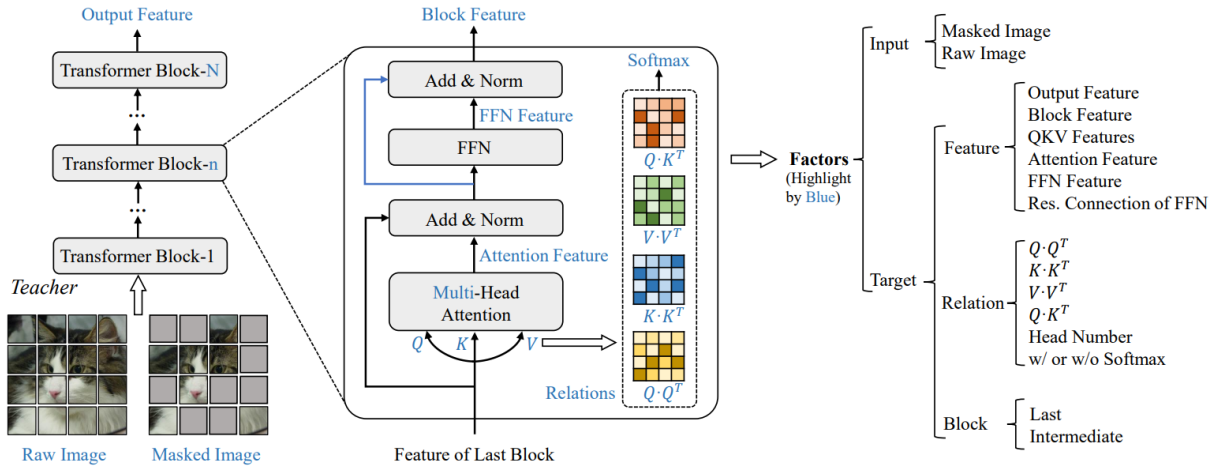


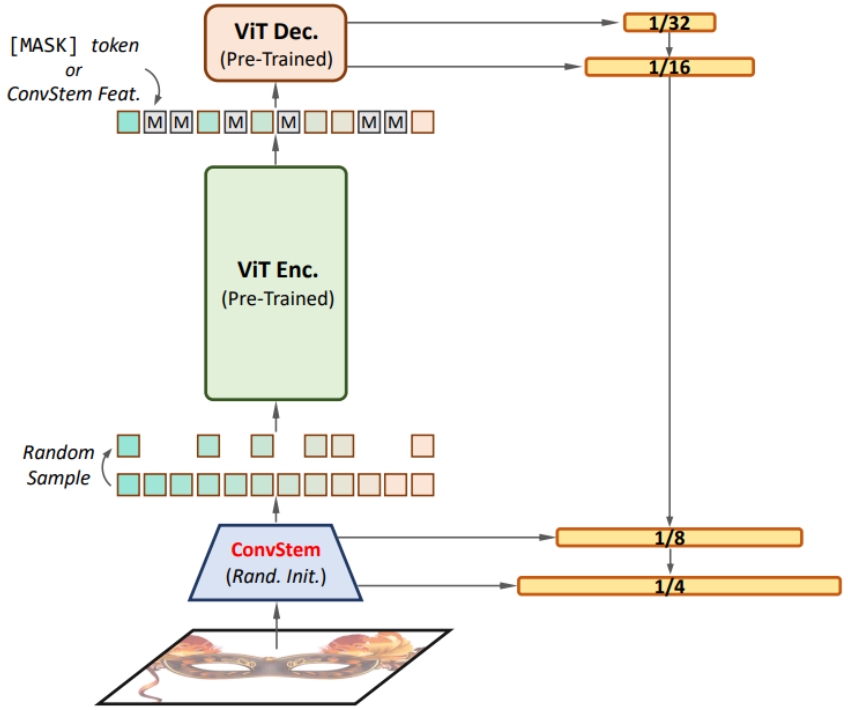
Figure 2. We comprehensively study a variety of factors (highlighted by Royal Blue) that may affect TinyMIM pre-training including input, distillation target (feature or relation) and target block.

Model	Param. (M)	Flops (G)	Top-1 (%)	mIoU
DeiT-T [44]	5.5	1.3	72.2	38.0
PVT-T [46]	13.0	1.9	75.1	39.8
CiT-T [39]	5.5	1.3	75.3	38.5
Swin [32]	8.8	1.2	76.9	40.4
EdgeViT-XS [35]	6.4	1.1	77.5	42.1
MobileViTv1-S [34]	4.9	2.0	78.4	42.7
MobileViTv3-S [45]	4.8	1.8	79.3	43.1
TinyMIM*-T (Ours)	5.8	1.3	**79.6**	**45.0**

Table 1. Comparison with state-of-the-art tiny Transformers with architecture variants. The parameters indicate the backbone parameter excluding the parameters of the last classification layer in classification or the decoder in segmentation. We report top-1 accuracy on ImageNet-1K classification and mIoU on ADE20K segmentation.

技术进展四：挖掘掩码图像建模的性质

- 对物体检测友好
  - ViTDet [Meta, ECCV 2022]
  - MIMDet [华科&腾讯，NeurIPS 2022]

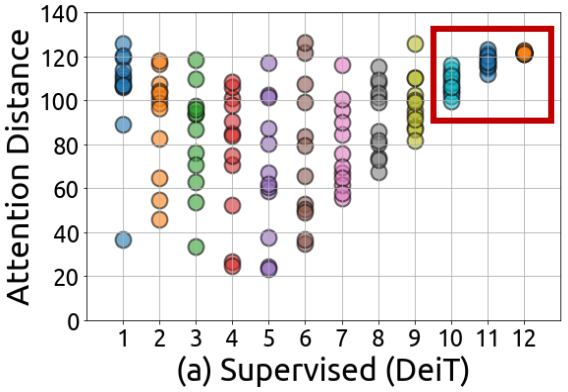


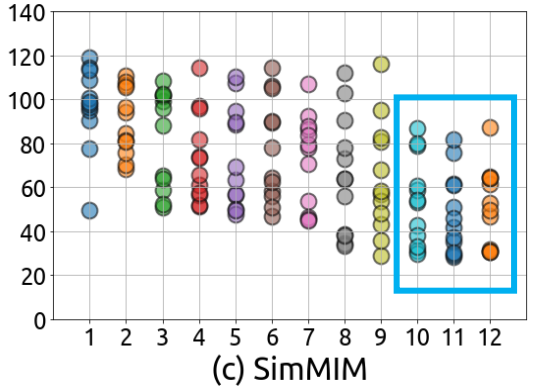
(b) Our MIMDet

技术进展四：挖掘掩码图像建模的性质

- 系统性研究掩码图像建模性质：MIM Dark Secrets [清华&中科大&微软, CVPR 2023]

不同层的注意力距离





语义理解任务

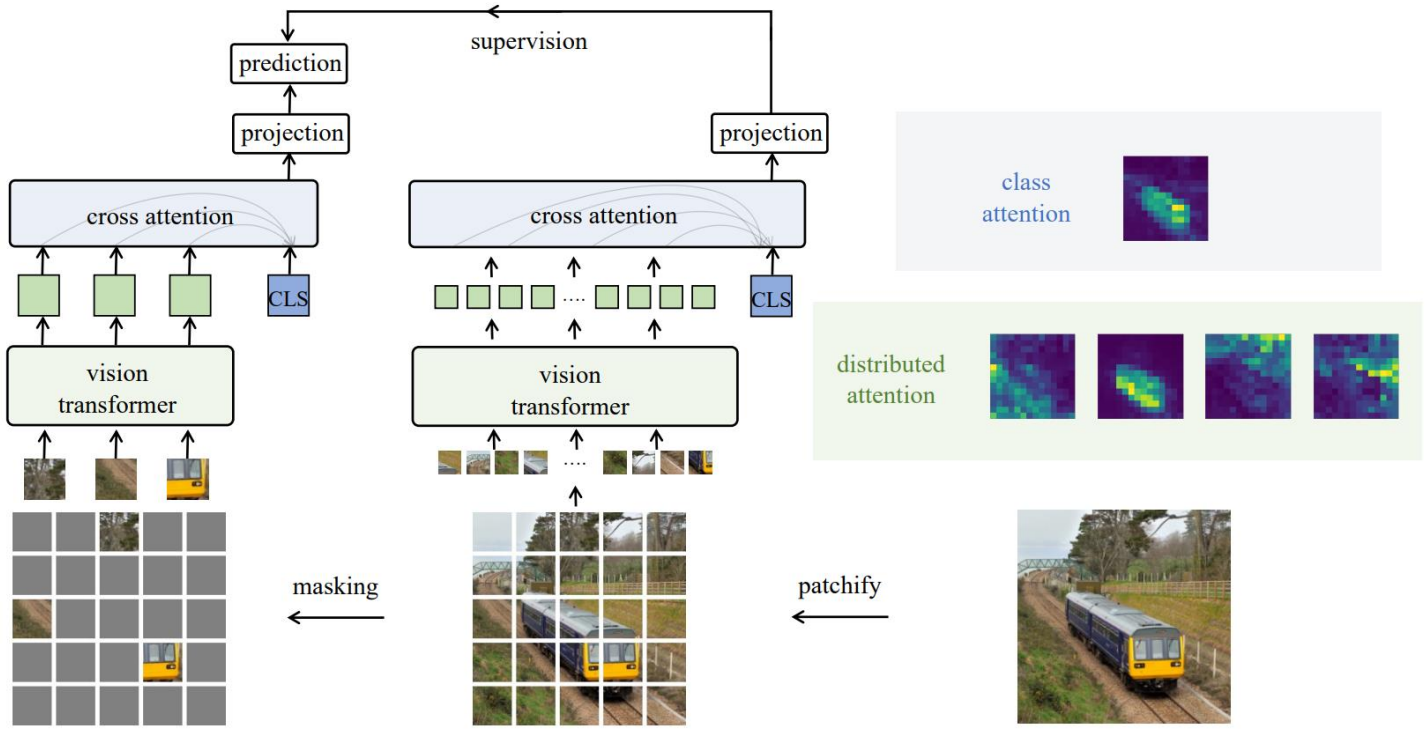
	Sup	MIM
K12-Set	**89.7**	86.1
Oxford Pets	**95.9**	90.9
Caltech101	**91.9**	85.5
SUN397	**72.3**	70.8

几何&运动任务

	Sup	MIM
Pose	75.9	**77.6**
Depth	0.335	**0.304**
Obj. Tracking	67.8	**70.0**

技术进展四：挖掘掩码图像建模的性质

- 掩码图像建模方法机制的其它解释
  - 效果好部分源自以随机掩码作为一种图像增强方法
    - ExtreMA [微软&CMU, Arxiv 2022]
    - 学习遮挡不变特征[旷视, Arxiv 2023]



ExtreMA框架

技术进展五：拓展到其它模态

- 拓展到视频
  - MaskFeat [Meta]
  - BEVT [复旦&微软]
  - VideoMAE [南大]
  - VideoMAE [Meta]
  - OmniMAE [Meta]
  - VideoMAE v2 [南大&上海AI Lab]

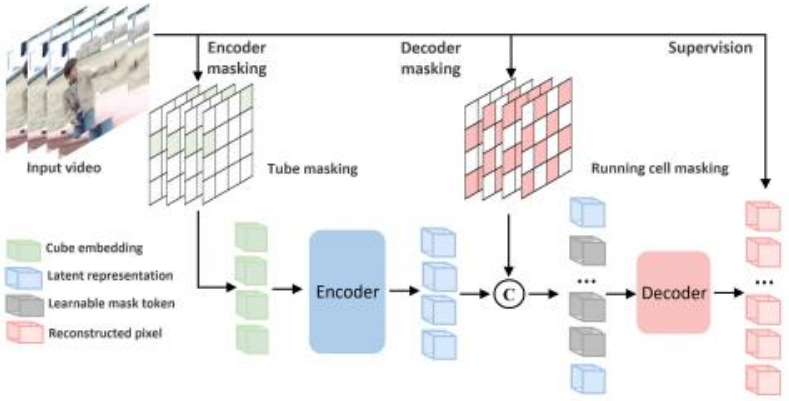


Figure 1. **VideoMAE with dual masking.** To improve the overall

(a) Kinetics 400

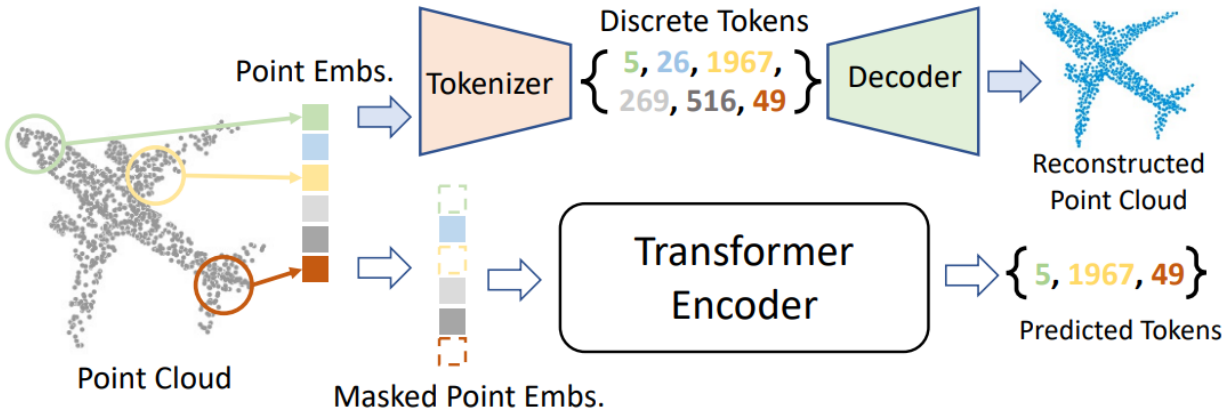
Method	Top 1	Top 5	Views	TFLOPs
I3D NL [74]	77.7	93.3	10 × 3	10.77
TDN [70]	79.4	94.4	10 × 3	5.94
SlowFast R101-NL [19]	79.8	93.9	10 × 3	7.02
TimeSformer-L [4]	80.7	94.7	1 × 3	7.14
MTV-B (320^2) [82]	82.4	95.2	4 × 3	11.16
Video Swin-L (384^2) [47]	84.9	96.7	10 × 5	105.35
ViViT-L FE [1]	81.7	93.8	1 × 3	11.94
MViTv2-L (312^2) [38]	86.1	97.0	40 × 3	42.42
MaskFeat [76]	87.0	97.4	4 × 3	45.48
MAE-ST [18]	86.8	97.2	4 × 3	25.05
VideoMAE [63]	86.6	97.1	5 × 3	17.88
VideoMAE V2-H	88.6	97.9	5 × 3	17.88
VideoMAE V2-g	88.5	98.1	5 × 3	38.16
VideoMAE V2-g (64×266^2)	**90.0**	**98.4**	2 × 3	160.30
Methods using in-house labeled data				
CoVeR (JFT-3B) [85]	87.2	–	1 × 3	-
MTV-H (WTS 280^2) [82]	89.9	98.3	4 × 3	73.57

(c) Something-Something V2

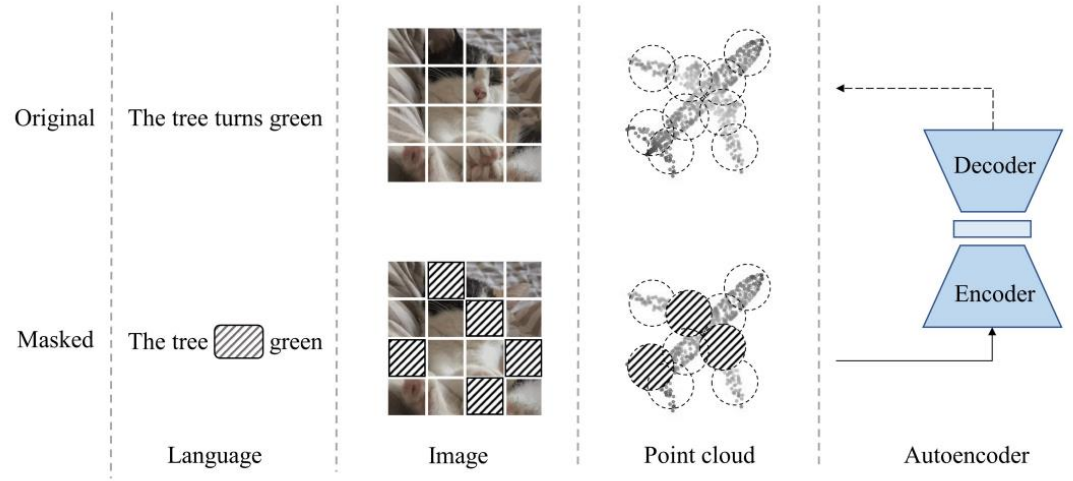
Method	Top 1	Top 5
SlowFast [19]	63.1	87.6
TEINet [46]	66.5	-
TEA [37]	65.1	89.9
TDN [70]	69.6	92.2
TimeSformer-L [4]	62.4	-
MFormer-HR [53]	68.1	91.2
ViViT-L FE [1]	65.9	89.9
Video Swin-B [47]	69.6	92.7
MViTv2-B [38]	72.1	93.4
MTV-B [82]	67.6	90.1
BEVT [72]	70.6	-
VIMPAC [60]	68.1	-
UniFormer [35]	71.2	92.8
MaskFeat [76]	75.0	95.0
MAE-ST [18]	75.5	95.0
VideoMAE [63]	75.4	95.2
VideoMAE V2-H	76.8	95.8
VideoMAE V2-g	**77.0**	**95.9**

技术进展五：拓展到其它模态

- 拓展到3D视觉中
  - PointBERT [清华&智源 CVPR2022]
  - Point-MAE [北大&新加坡国立&鹏程实验室 ECCV2022]



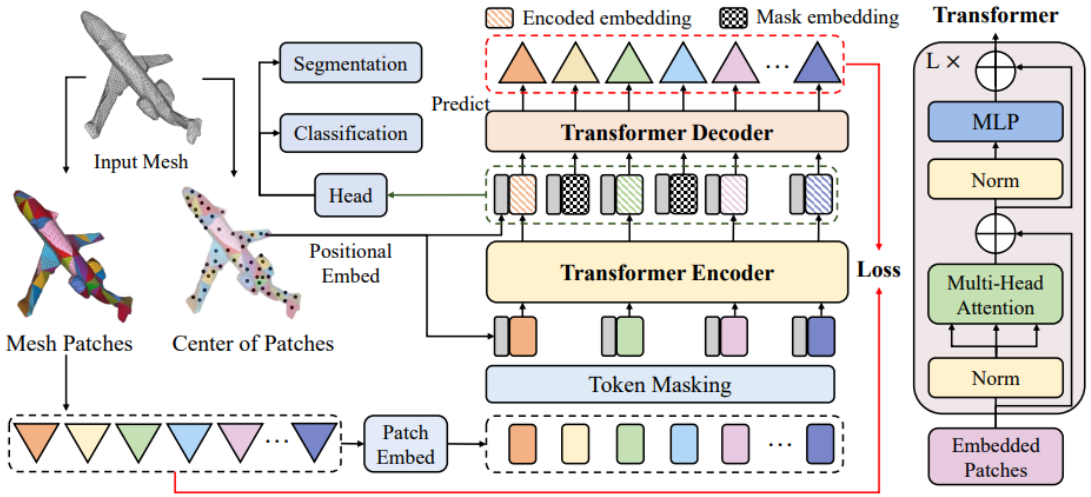
PointBERT [清华&智源 CVPR2022]



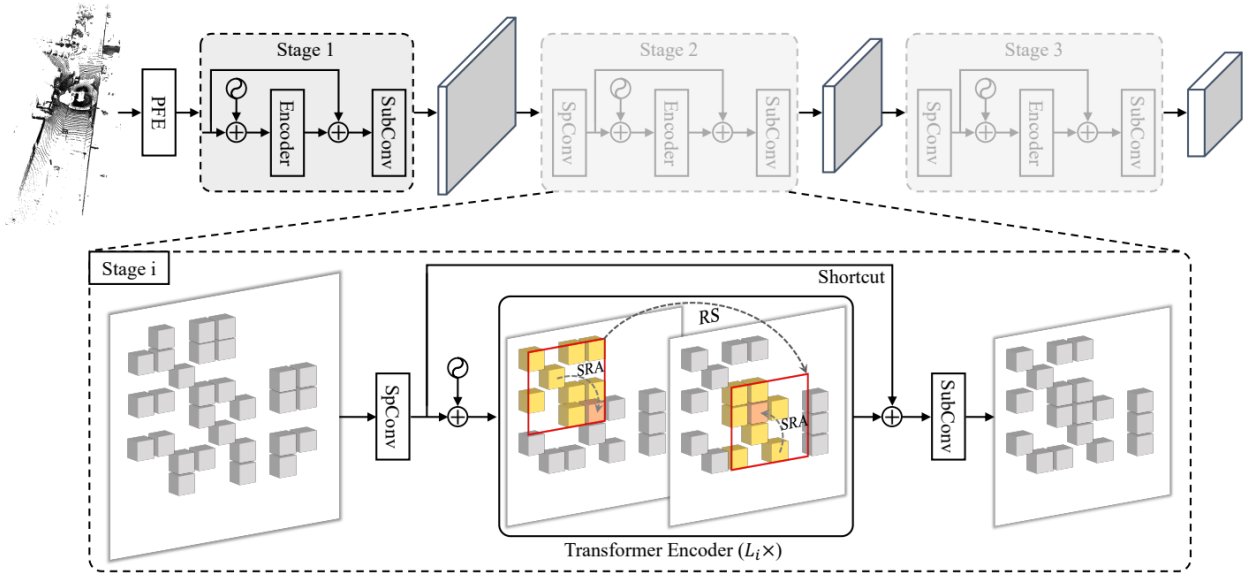
Point-MAE [北大&新加坡国立&鹏程实验室 ECCV2022]

技术进展五：拓展到其它模态

- 拓展到3D视觉中
  - MeshMAE [武大&京东 ECCV2022]
  - GD-MAE [浙大&上海AI Lab CVPR2023]



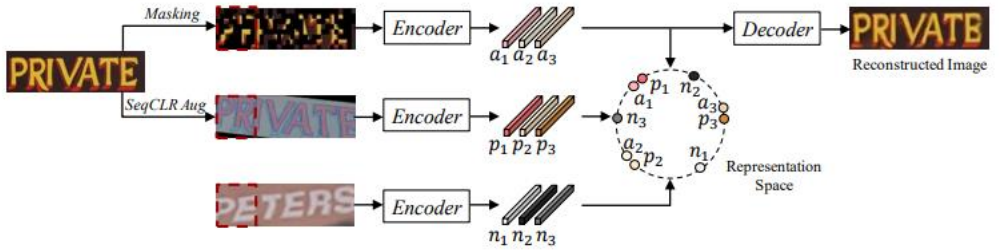
MeshMAE [武大&京东 ECCV2022]



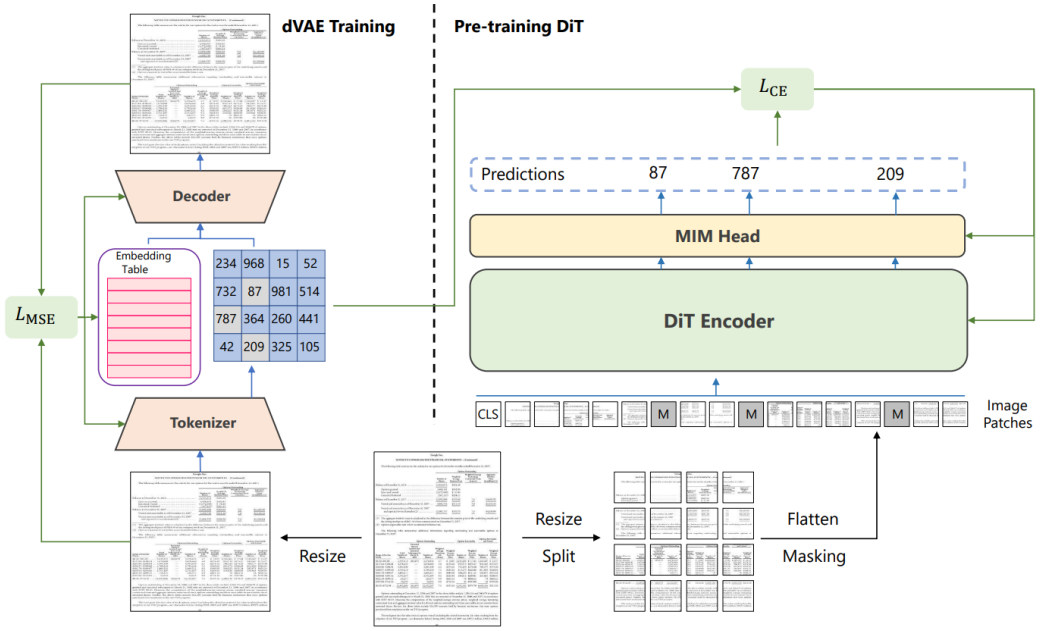
GD-MAE [浙大&上海AI Lab CVPR2023]

技术进展五：拓展到其它模态

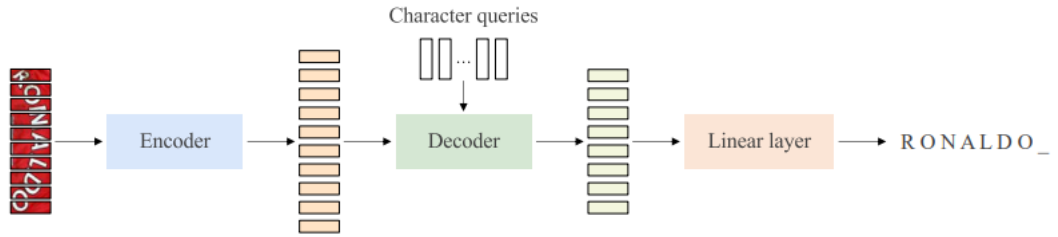
- 拓展到OCR
  - GiD [华科&华为 Arxiv2022]
  - MaskOCR [百度 Arxiv2022]
  - DiT [上交&微软 Arxiv2022]



GiD [华科&华为 Arxiv2022]



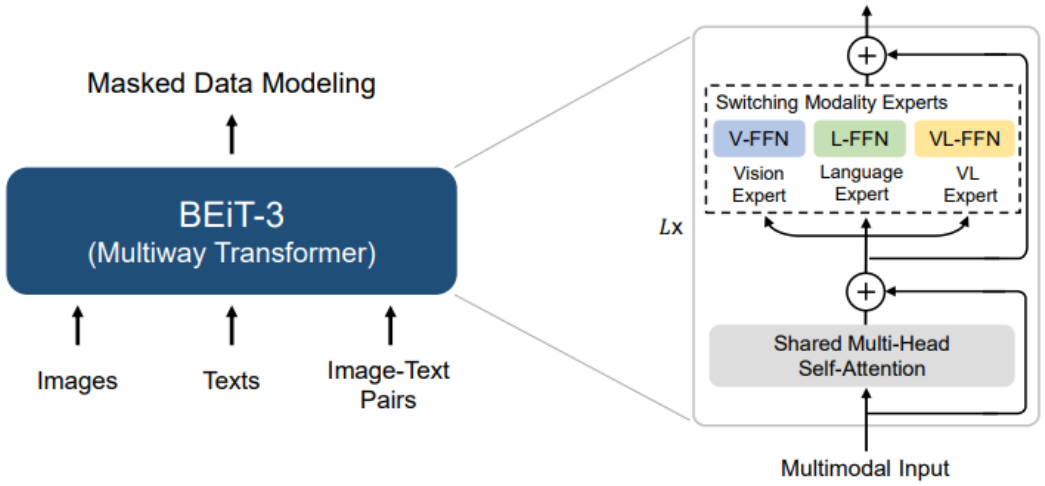
DiT [上交&微软 Arxiv2022]

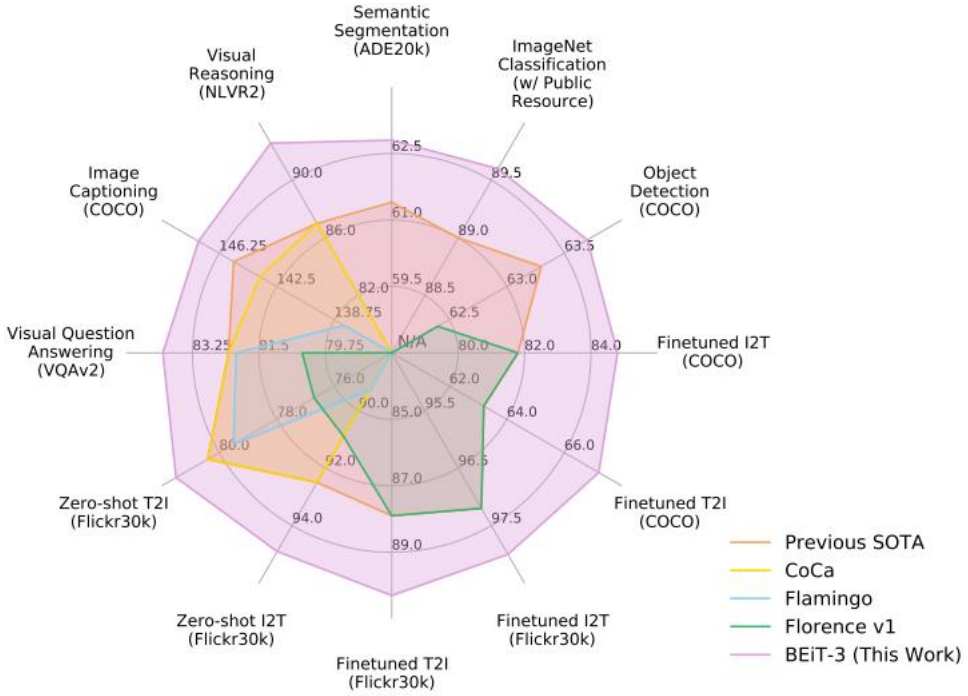


MaskOCR [百度 Arxiv2022]

技术进展五：拓展到其它模态

- 多模态中的掩码建模
  - BEiT-3 [微软 Arxiv2022]





小结：自监督学习年度进展——掩码图像建模

- 技术进展趋势一：掩码图像建模的**改进**
  - 以CLIP模型特征为重构目标: MVP[中科大&华为] BEiT-V2 [微软] MILAN [普林斯顿&阿里] 等
  - 深度监督：DeepMIM [微软]
- 技术进展趋势二：发现掩码图像建模对**大模型**比较友好
  - 信息量是关键：Swin V2[中科大&微软, CVPR2022] CAE v2 [百度] EVA [智源] InternImage [商汤]等
  - 大数据表现如何？Data Scaling [清华&微软 CVPR22] MAE on Billion-data [Meta Arxiv2023]
- 技术进展趋势三：针对**小模型**的掩码图像建模训练
  - 主要思想——蒸馏：TinyMIM [微软, CVPR2023] Lightweight MAE [自动化所，Arxiv2023]
- 技术进展趋势四：挖掘掩码图像建模的**性质**
  - 对物体检测友好：ViTDet [Meta, ECCV 2022] MIMDet [华科&腾讯，NeurIPS 2022]
  - 性质的系统研究：MIM Dark Secrets [清华&中科大&微软, CVPR 2023]
  - 随机掩码作为一种图像增强方法：ExtreMA [微软&CMU, Arxiv 2022] 学习遮挡不变特征[旷视, Arxiv 2023]
- 技术进展趋势五：**拓展到其它模态**
  - 视频/3D/OCR/多模态等

未来展望：生成式预训练

If I cannot create, I don’t understand.

——Richard Feynman

谢谢大家！