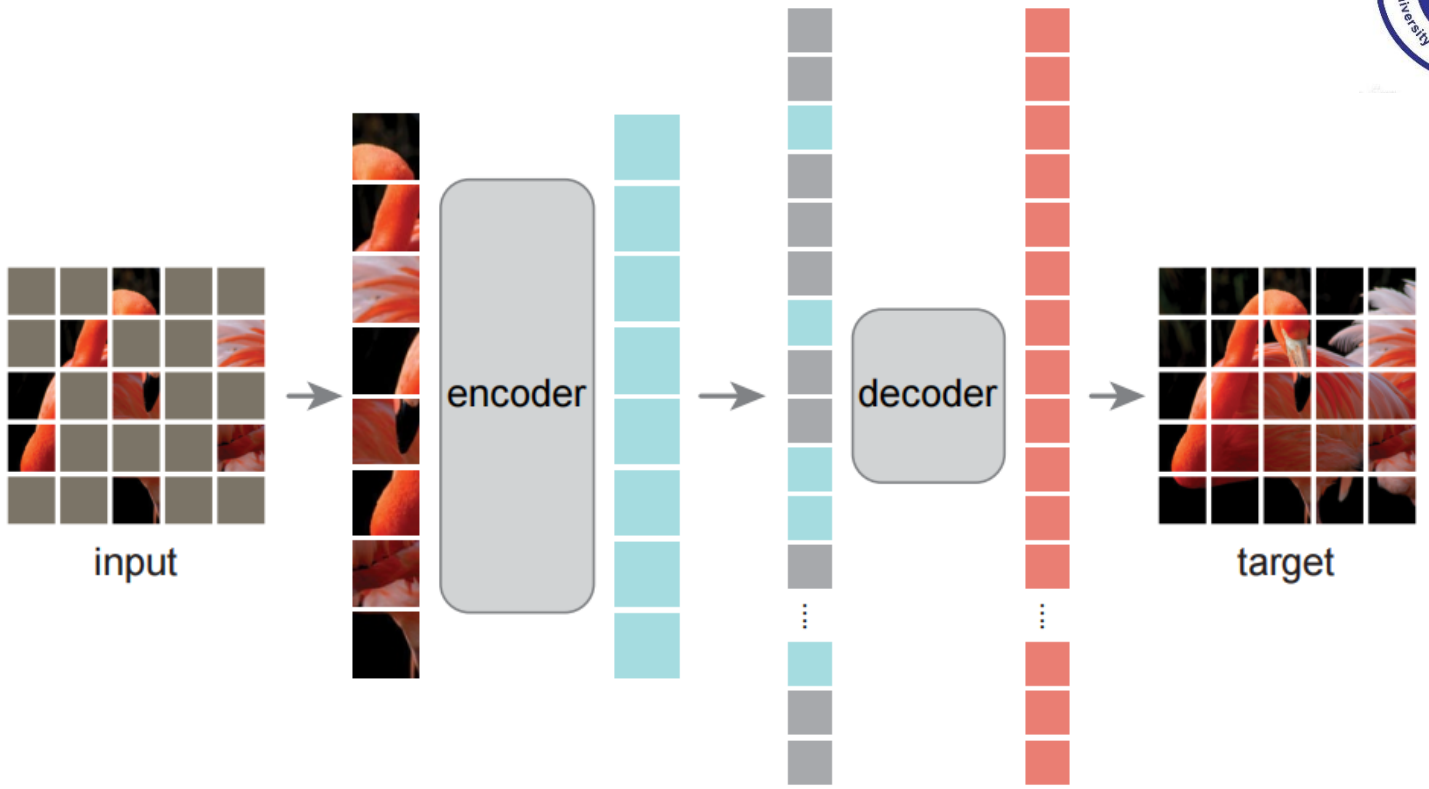
Efficient Masked Autoencoders with Self-Consistency

Zhaowen Li, Yousong Zhu*, Zhiyang Chen, Wei Li, Rui Zhao, Chaoyang Zhao, Ming Tang, *Member, IEEE*
Jinqiao Wang*, *Member, IEEE*,

Conference: PAMI **Year: 2024**

MAE

- 将图像划分为 patch（如16×16），随机遮盖其中75%。
- 编码器仅处理未遮盖的patch，提取语义特征。
- 解码器基于编码结果重建被遮盖的图像块。



研究背景与动机

MAE目前还有什么问题:

1、数据利用率低:

MAE 每次只输入 25% patch 到编码器

75% 的图像信息未被编码，学习效率低下

2、重建语义不一致:

不同掩码下对同一位置预测差异大

会导致语义漂移，影响迁移性能

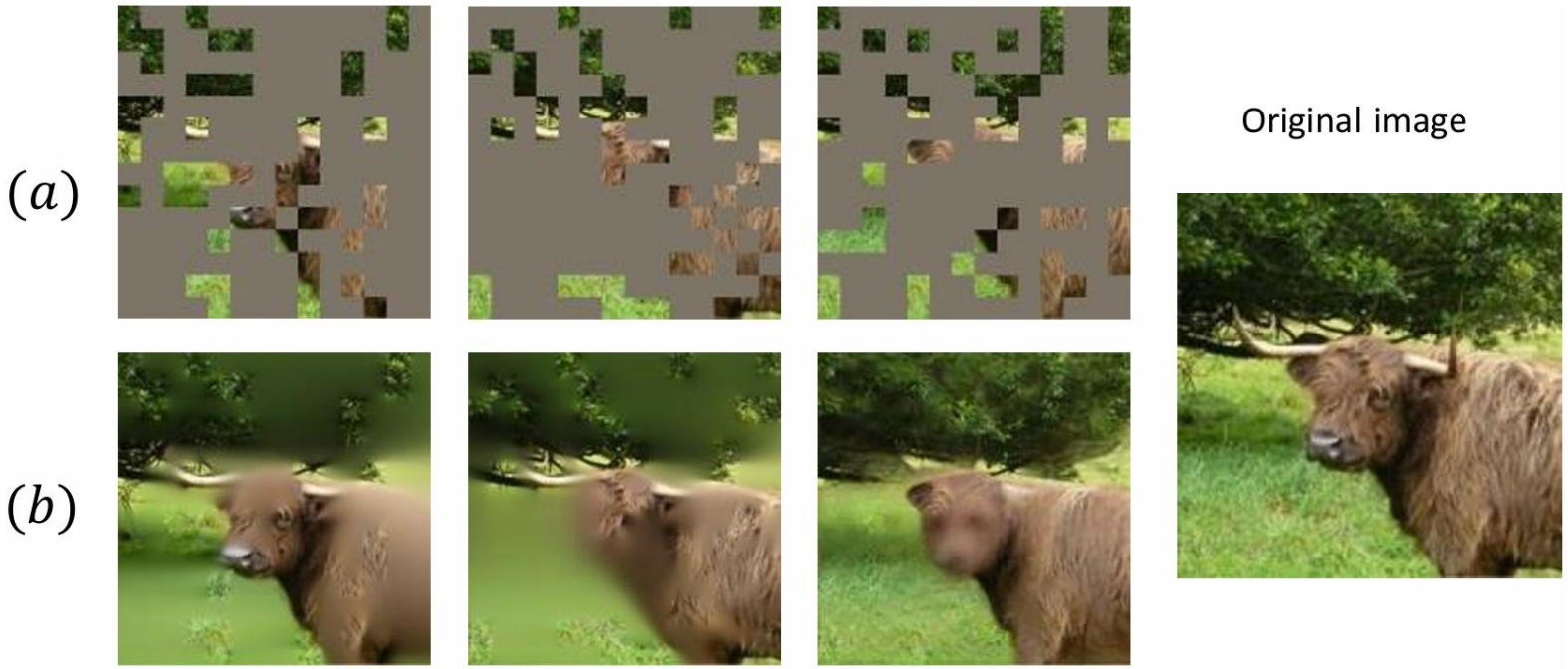


图 1. MAE [21] 在不同掩码种子下的重建结果不同。(a) 不同掩码种子采样得到的可见图像块组合。(b) MAE 的重建结果。图 (b) 中的三个重建结果中，仅第一个是正常的牛，第三个甚至重建出了狗。这些由 MAE 重建的图像语义不一致。

EMAE 整体思路

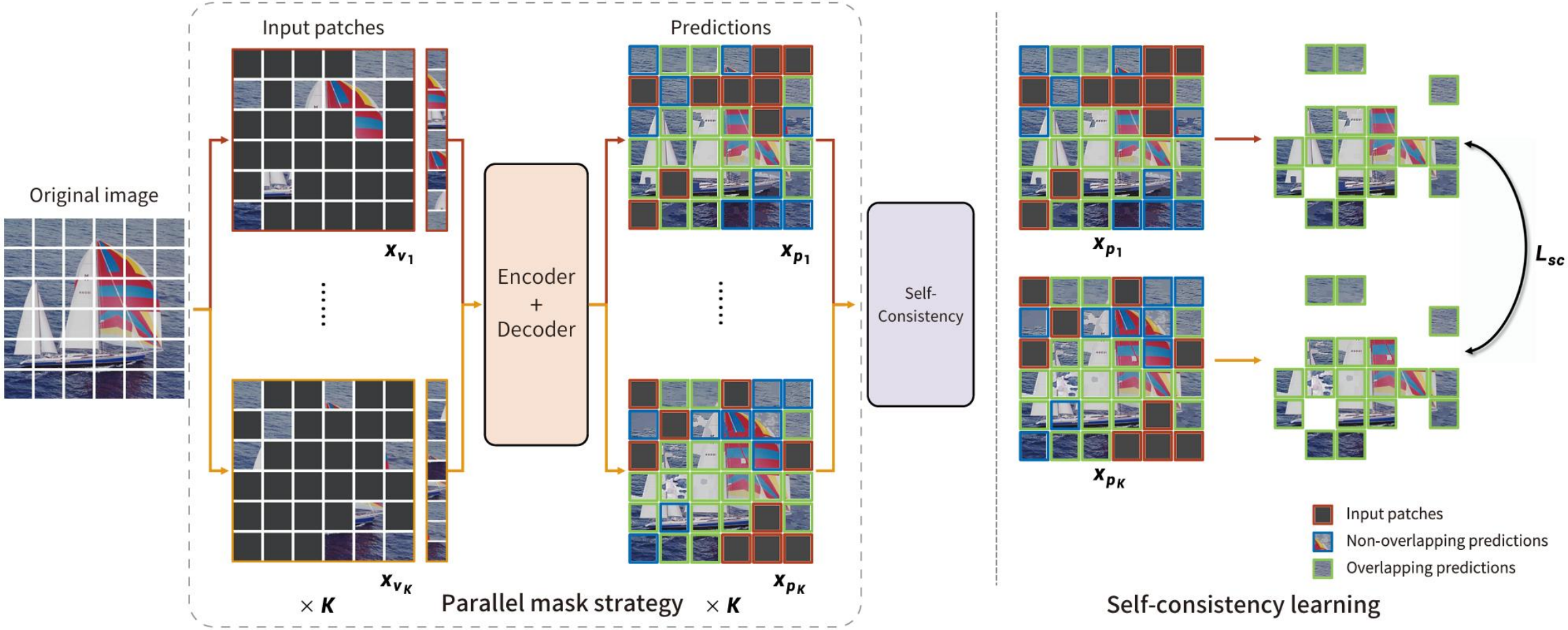
EMAE 提出两个核心设计：

1、并行遮盖机制：

提升每轮训练中图像信息的利用率

2、一致性建模机制：

增强对同一 patch 的语义稳定性



并行遮盖策略

图 2. EMAE 方法示意图。整张图像首先被划分为 N 个图像块，然后并行策略将这 N 个图像块划分为 K 个不重叠的部分 $x_v^1,\ldots,x_v^K$，每一部分大小相同，每部分包含 N/K 个随机且不重叠的可见图像块。随后，每一部分被输入到编码器-解码器结构中执行 MIM 任务，生成 $x_m^1,\ldots,x_m^K$。此外，引入自一致性学习以将相同位置的重叠预测拉近一致。

Algorithm 1 Pseudocode of mask generation of the parallel mask strategy in a PyTorch-like style.

```
# x: the input image
# K: the number of non-overlapping parts

# map an image into multiple image patches
x = patchify(x)
N, D = x.shape # length, dim

tensor = rand(N) # tensor in [0, 1]

# sort the tensor in ascending order
ids = argsort(tensor)

# acquire the position of each element
ids_tensor = argsort(ids)

# divide the whole data into K parts
for i in range(1,K+1):

   # obtain the i-th visible patches
   ids_i = ids[(i-1)*(N/K):i*(N/K)]
   x_v_i = gather(x, dim=0, index=ids_i)

   # obtain the i-th mask
   m_i = ones(N)
   m_i[(i-1)*(N/K):i*(N/K)] = 0
   m_i = gather(m_i, dim=0, index=ids_tensor)

   # obtain the i-th masked patches
   x_m_i = x[m_i].reshape(-1,D)
```

`rand`: returns a tensor filled with random numbers from a uniform distribution on the interval [0,1];
`gather`: gathers values along an axis specified by dim;
`argsort`: returns the indices that sort a tensor along a given dimension in ascending order by value.

一致性建模机制

1、解决MAE预测不一致性问题：

不同遮盖上下文预测不同

2、对同一patch在不同部分的预测，

约束其特征相似

3、基于并行遮盖的单次前向传播，

联合优化MIM损失和一致性损失

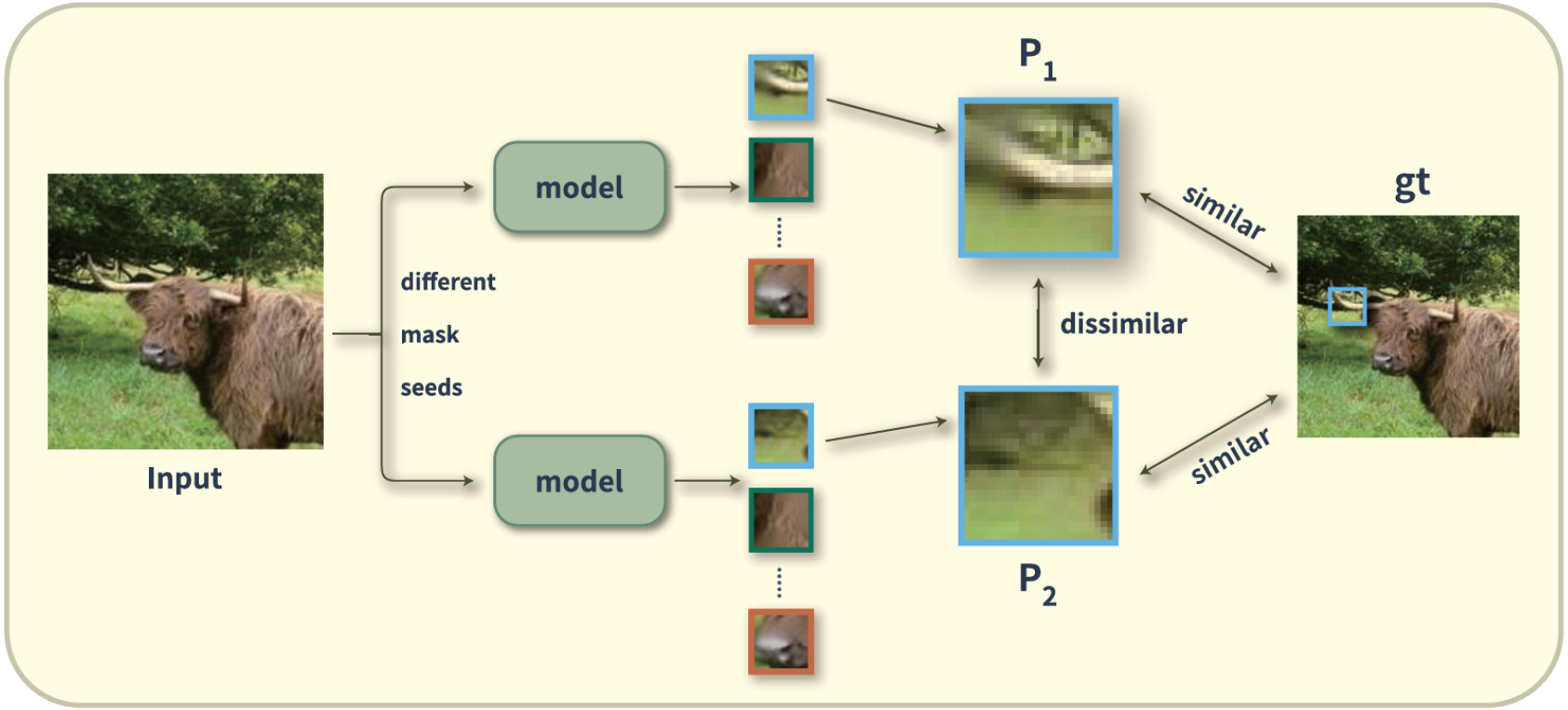


图 2. MIM 预训练模型在不同掩码种子下对相同位置的不同预测结果。P1 和 P2 是相同位置的预测，gt 是该位置的真实图像块。虽然 P1 和 P2 分别与 gt 相似，但二者之间的语义却不一致。

$$\mathcal{L}_{total}(\mathbf{x}) = \mathcal{L}_{whole}(\mathbf{x}) + \mathcal{L}_{consistency}(\mathbf{x}),$$

TOTAL SPEEDUP FOR DIFFERENT METHODS ON 64 NVIDIA A100 GPUS.

Encoder	Architecture	Total iterations	Total FLOPs per iteration	Total times (hours)	Total speedup
MAE	ViT-L	1×	1×	180.76	1×
Supervised	ViT-L	0.125×	∼ 4×	23.83	**7.58**×
EMAE	ViT-L	0.125×	∼ 4×	23.75	**7.61**×

表 1. EMAE 通过并行掩码策略利用完整图像，提高每轮训练的信息量，因此可在更少的迭代中收敛，显著缩短训练时间。在实际对比中，仅需 13% 的 MAE 训练时间，EMAE 即可达到或超越 MAE 的最终性能。此外，EMAE 能充分利用 GPU 并行性，避免了串行计算带来的时间浪费，进一步提升效率。

实验

ACCURACY ON THE IMAGENET VAL SET.

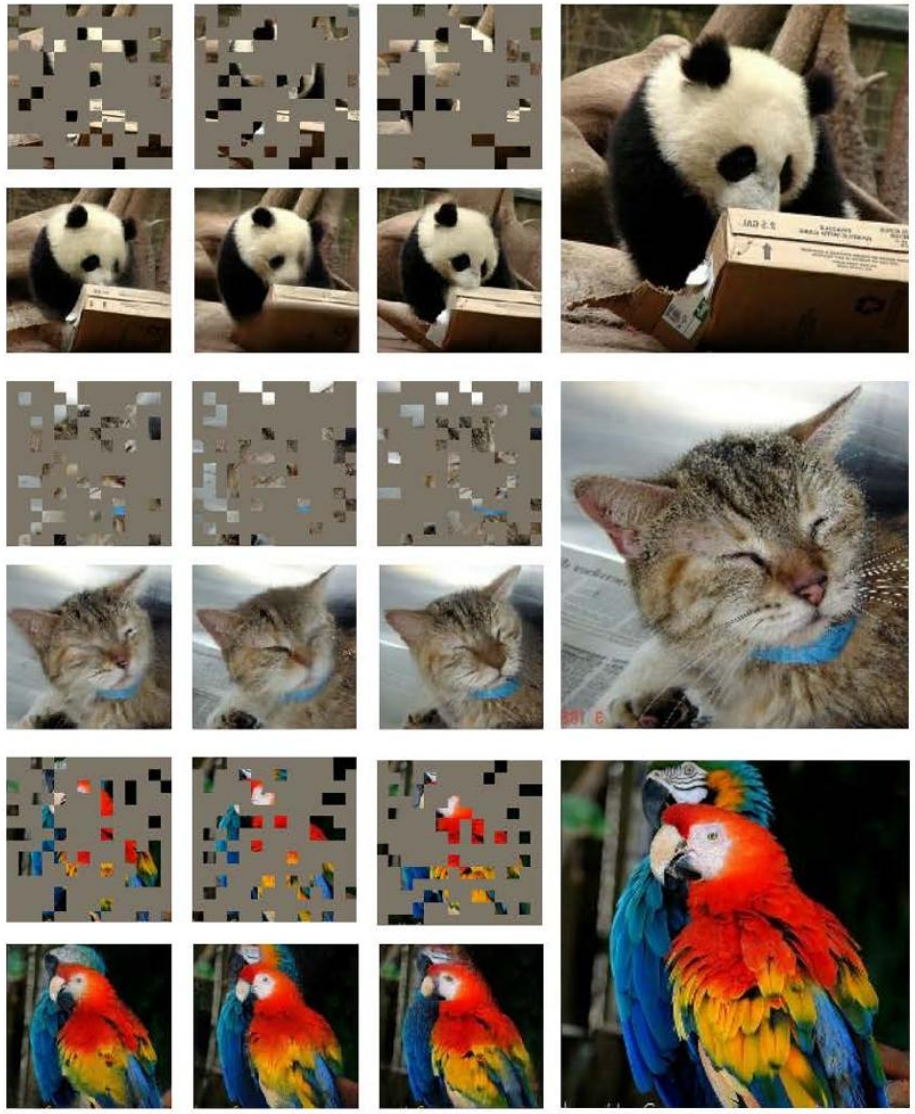
Method	Architecture	Pre-train epochs	Linear Probing Top-1 Acc	Fine-tuning Top-1 Acc
Supervised learning on ImageNet:				
scratch [71]	ViT-S	300	-	79.9%
scratch [21]	ViT-B	300	-	82.3%
scratch [21]	ViT-L	300	-	82.6%
Contrastive learning:				
MoCo v3 [32]	ViT-B	300	76.2%	83.2%
DINO [46]	ViT-B	400	**78.2%**	**83.4%**
MSN [42]	ViT-B	600	-	**83.4%**
Masked image modeling + contrastive learning:				
MST [17]	ViT-S	100	75.0%	-
AttMask [75]	ViT-S	100	76.1%	81.3%
iBOT [18]	ViT-S	100	74.4%	81.1%
iBOT [18]	ViT-S	3,200	77.9%	82.3%
iBOT [18]	ViT-B	1,600	79.5%	84.0%
iBOT [18]	ViT-L	1,200	**81.0%**	**84.8%**
Masked image modeling:				
CAE [71]	ViT-S	300	50.8%	81.8%
CAE [71]	ViT-B	800	68.3%	83.6%
BEiT [19]	ViT-B	800	56.7%	83.2%
BEiT [19]	ViT-L	800	73.5%	85.2%
SimMIM [20]	ViT-B	800	56.7%	83.8%
data2vec [44]	ViT-B	800	-	84.2%
data2vec [44]	ViT-L	1600	77.3%	86.6%
I-JEPA [45]	ViT-B	600	72.9%	-
I-JEPA [45]	ViT-L	600	**77.5%**	-
MAE [21]	ViT-B	300	61.5%	82.9%
MAE [21]	ViT-B	800	64.4%	83.4%
MAE [21]	ViT-B	1600	67.8%	83.6%
MAE [21]	ViT-B	2400	68.2%	83.8%
MAE [21]	ViT-L	1600	75.6%	85.9%
MAE [21]	ViT-H	1600	76.6%	**86.9%**
EMAE	ViT-B	300	68.2%	83.8%
EMAE	ViT-B	800	70.4%	84.0%
EMAE	ViT-L	800	**76.7%**	**86.3%**

主流自监督学习方法比较。EMAE 在 ImageNet 训练集上预训练，并在性能上超过了以往的 MIM 方法。我们在两个监督训练设置下测试预训练模型的性能：1）线性探测；2）端到端微调。表中报告了 ImageNet 验证集上的 Top-1 准确率。

Method	Pre-train epochs	Pre-train data	Object detection			Instance segmentation		
			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Supervised [71]	300	ImageNet-1K	47.9%	-	-	42.9%	-	-
MoCo v3 [32]	600	ImageNet-1K	47.9%	-	-	42.7%	-	-
BEiT [19]	800	ImageNet-1K + DALLE	49.8%	-	-	44.4%	-	-
CAE† [71]	1600	ImageNet-1K	50.0%	70.9%	54.8%	44.0%	67.9%	47.6%
MAE [21]	1600	ImageNet-1K	50.4%	70.8%	55.7%	44.9%	68.3%	48.9%
EMAE	300	ImageNet-1K	50.6%	70.9%	56.1%	45.0%	68.6%	49.3%
EMAE	800	ImageNet-1K	**51.4%**	**72.2%**	**56.5%**	**45.7%**	**69.4%**	**49.8%**

在 COCO 数据集上使用 Mask R-CNN 进行目标检测与实例分割的结果。所有方法均使用 ViT-B 架构，采用 FPN 的 Mask R-CNN 并报告 COCO val2017 上的边界框 AP 和掩码 AP。EMAE 的表现优于之前的主流自监督学习方法。

Method \ Pre-epochs	100	300	800
(a) MAE mask ratio = $\frac{3}{4}$	54.8%	61.5%	64.4%
(b) EMAE K = 4	60.9%	68.4%	70.4%
(c) MAE mask ratio = $\frac{6}{7}$	53.3%	61.0%	63.9%
(d) EMAE K = 7	60.5%	66.5%	68.1%
(e) MAE mask ratio = $\frac{13}{14}$	46.4%	54.7%	60.7%
(f) EMAE K = 14	52.5%	61.0%	66.7%



EMAE 在不同掩码种子下的重建结果。使用不同随机种子从同一图像中采样出不同的可见块组合，并将其输入 EMAE 进行重建，图中展示了三个重建示例。这些重建结果语义相似，并相互接近，验证了我们提出的自一致性学习的有效性。

Method	Pre-train data	Pre-train epochs	Top-1
(a) baseline	ImageNet-1K	200	58.8%
(b) pure random mask			62.5%
(c) parallel mask			63.4%
(d) complementary mask			46.7%

并行掩码在相同训练周期下能提升性能

Method	Pre-train data	Pre-train epochs	Top-1
(a) pixel reconstruction	ImageNet-1K	200	63.4%
(b) pure self-consistency			64.9%
(c) pixel reconstruction + self-consistency			65.0%

自一致性损失能提升性能

结论

EMAE提出了两项关键创新:

首先，通过并行掩码策略，将数据利用率提升至100%，大幅缩短预训练时间;

其次，引入自一致性学习，显著增强了掩码图像建模（MIM）的预测一致性和可靠性。

实验验证表明，EMAE在ImageNet、COCO等数据集上的多种下游任务中展现了卓越的泛化能力，性能优于MAE。

Salience-Based Adaptive Masking: Revisiting Token Dynamics for Enhanced Pre-training

Hyesong Choi[1], Hyejin Park[1], Kwang Moo Yi[2],
Sungmin Cha[3], and Dongbo Min[1]

[1] Ewha W. University
[2] University of British Columbia
[3] New York University

Conference: ECCV Year: 2024

研究背景

MAE目前还有什么问题:

1、随机掩码忽略token重要性:

掩盖背景而非关键语义区域（如物体核心）。

2、掩码比率敏感

微调（如0.75→0.6）导致性能大幅波动（如MAE准确率降2-3%）。

3. 很多基于注意力的方法，

会带来**大量的额外计算开销**

而且有些鲁棒性不足，性能随比率变化波动。

动机

本文的创新点：

1、基于显著性掩码（SBAM）：

利用注意力机制的outgoing weight，优先掩码语义贡献大的token（如物体核心）。

2、自适应掩码比率（AMR）：

根据token显著性分布动态调整比率。例：熊特写需高比率，远景飞机需低比率。

本文目标：

提升MIM预训练效率，降低比率敏感性。

SBAM（显著性自适应掩码）

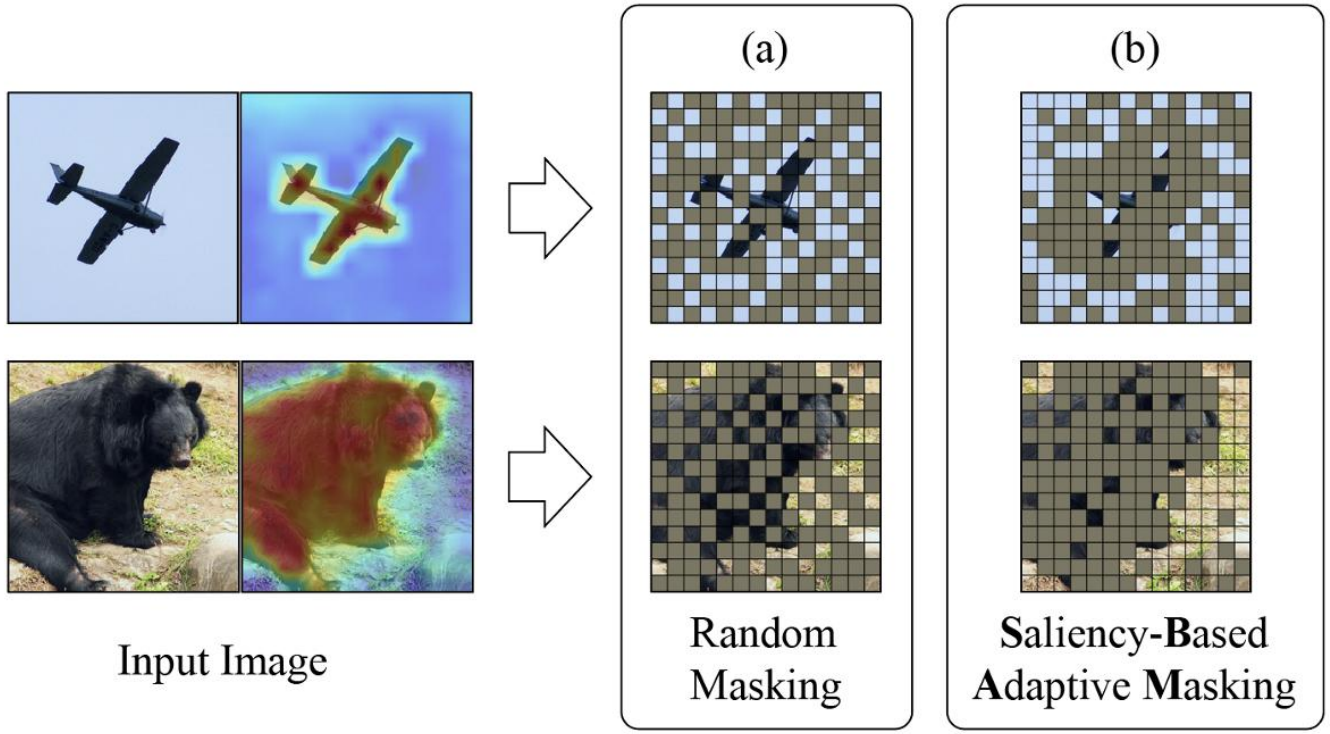
1.SBAM： 显著性自适应掩码，解决随机掩码忽略token重要性问题。

2.显著性计算： 利用注意力机制，优先掩码语义关键区域（如物体核心）。

3.随机性引入： 添加均匀分布噪声，增加掩码多样性。

4.流程： 提取token序列 → 计算显著性分数 → 生成自适应掩码。

5.目标： 提升预训练效率和语义理解能力。



随即掩码与SBAM掩码策略的区别

SBAM具体实现

输入token序列： $X \in \mathbb{R}^{N \times L \times D}$

计算注意力图（$A_{n,i,j}$：第n张图中，第i个token对第j个token的相似度得分）：

$$A = X \cdot X^{\top} \in \mathbb{R}^{N \times L \times L}$$

归一化为注意力概率矩阵：

$$\hat{A}_{n,i,j} = \frac{e^{a_{n,i,j}}}{\sum_k e^{a_{n,i,k}}}$$

显著性向量（表示每个token对其他token的重要性）：

$$S = \mathcal{N}\left(\sum_{j=1}^{L} \hat{A}_{:,j,:}\right)$$

$$\mathcal{N}(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$$

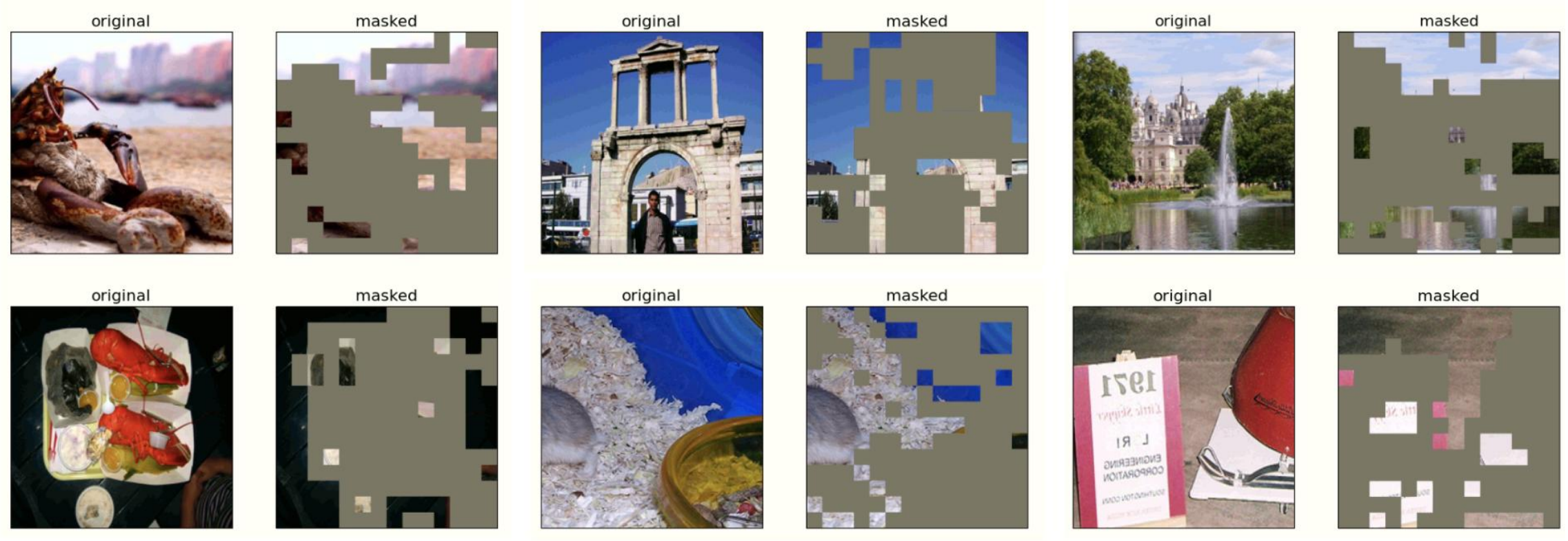
加入随机扰动（提升多样性）：

$$\tilde{S} = S + \mathcal{U}([0, 0.5))$$

对每张图显著性排序，选前K个token进行mask（设K = $\lceil L \cdot \gamma \rceil$）

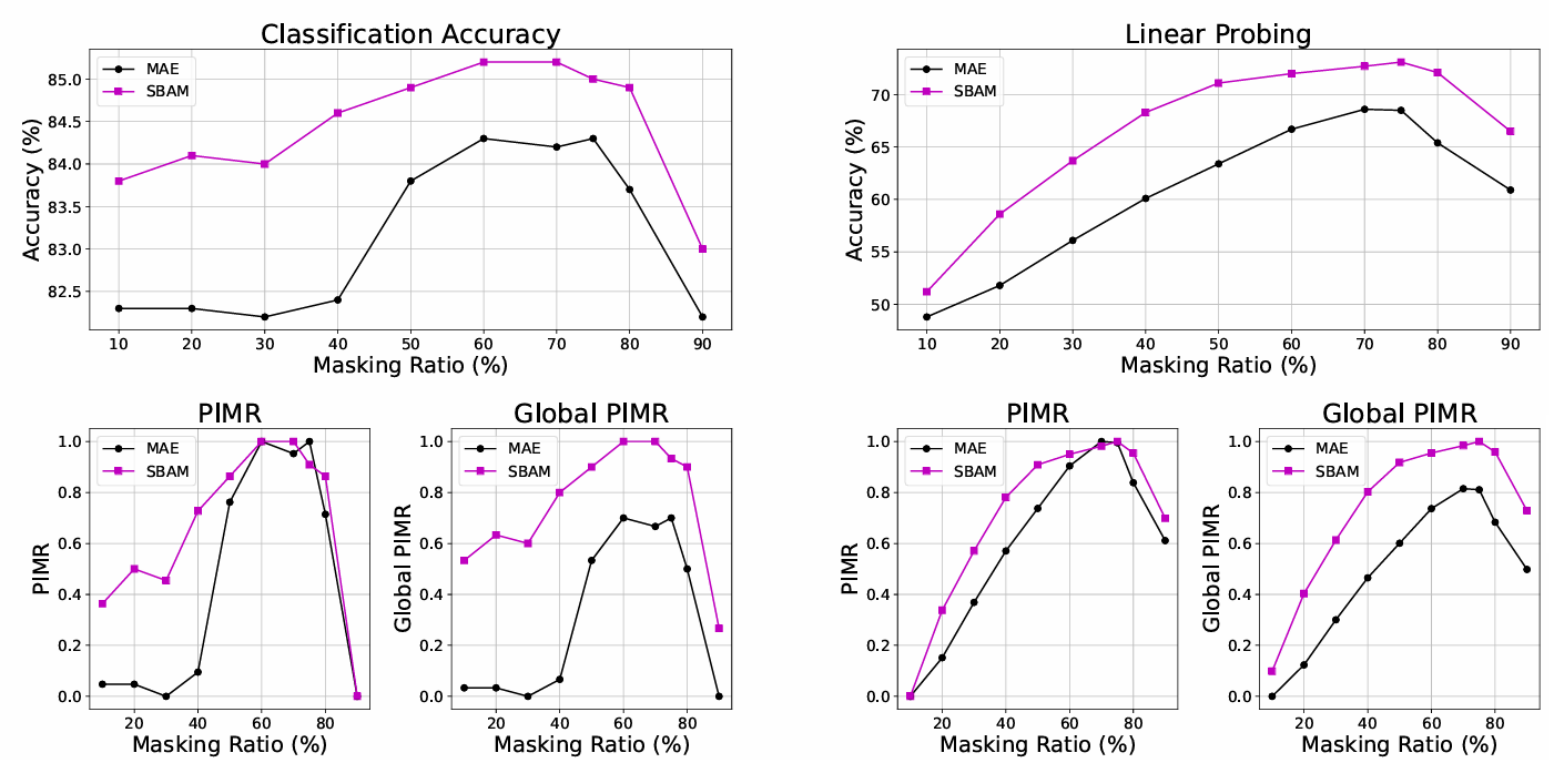
构造二值掩码矩阵 $M \in \{0,1\}^{N \times L}$

SBAM效果



SBAM 示例。SBAM 引入了“token 显著性”以优先并掩码高重要性的 token。因此，从定性上确认了图像中贡献较大的重要对象被选择性地掩码。此外，通过将随机性与 token 显著性结合，掩码以概率方式分配给背景和不太重要的 token，从而丰富了 token 掩码的多样性。

SBAM效果

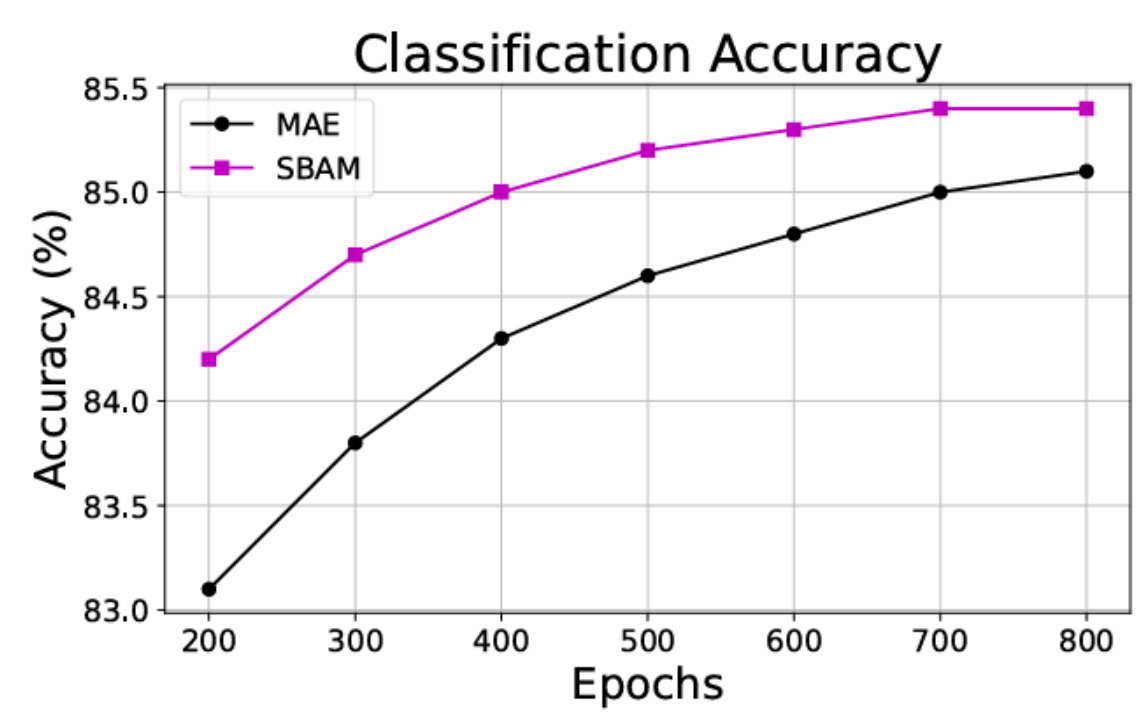


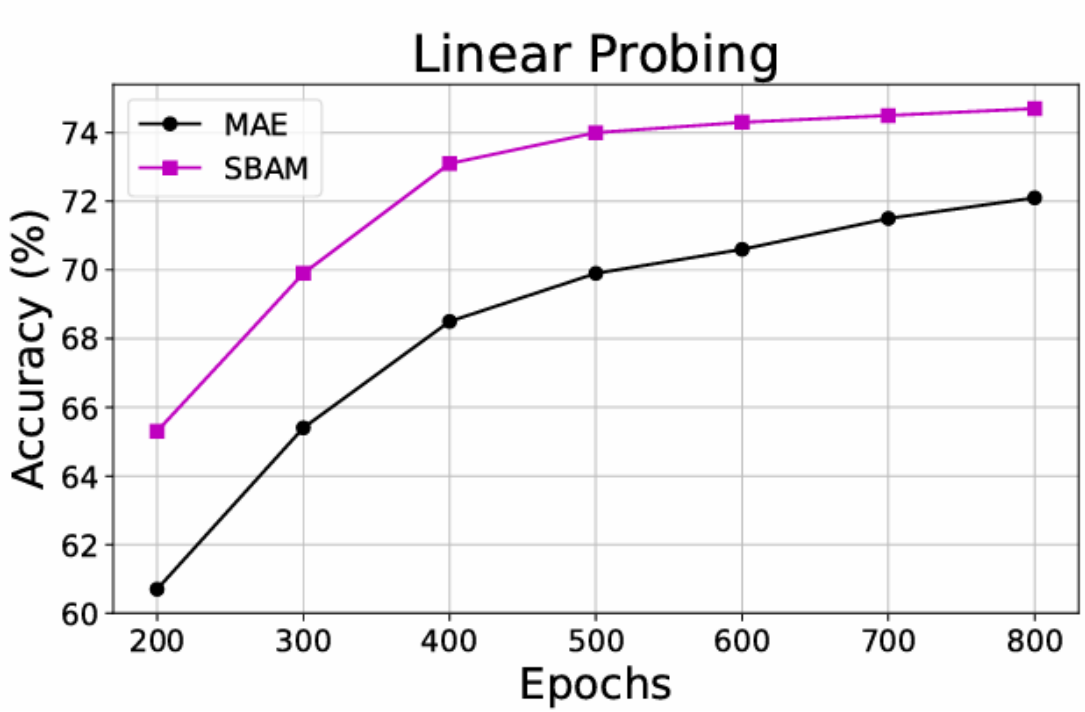
$$\text{PIMR}(M) = \frac{P(M) - P(M_{min})}{P(M_{max}) - P(M_{min})},$$

$$\text{Global PIMR}(M) = \frac{P(M) - P(M_{Gmin})}{P(M_{Gmax}) - P(M_{Gmin})},$$

为了评估 SBAM 的稳健性，我们对比分析了其在 ImageNet-1K 数据集上的图像分类性能，并与基线方法 MAE 进行了比较，采用 ViT-L 作为骨干网络。上图展示了在不同掩码比例下的方法性能，下图展示了掩码比例提升带来的性能提升（PIMR）和全局 PIMR。这些指标衡量了模型在从最低到较高掩码比例下的性能提升程度。SBAM 在所有指标中均显著优于 MAE，展示出其在处理多种掩码比例方面的卓越效果以及增强的预训练性能。

SBAM效果





曲线表明，SBAM 在每一个训练轮数上都优于 MAE，并且验证了其更快达到收敛性能水平。

AMR（自适应掩码比率策略）

1、动机：不同图像需要不同遮挡强度

熊特写图：目标区域密集 → 应遮更多

远景飞机图：大量背景 → 应遮更少

而固定75% mask比例无法自适应，影响效果。

2、基于显著性分布动态调整掩码比率：

利用token显著性S判断图像该遮掩多少。例：熊特写需高比率，远景飞机需低比率。

3、做法：

$$R_{dyna} = r - \Delta r + 2\Delta r \times \mathrm{mean}(1_{S>\delta}).$$

r: 基础掩码比率

Δr: 比率变化范围

S: 显著性分数

δ: 显著性阈值

AMR（自适应掩码比率策略）

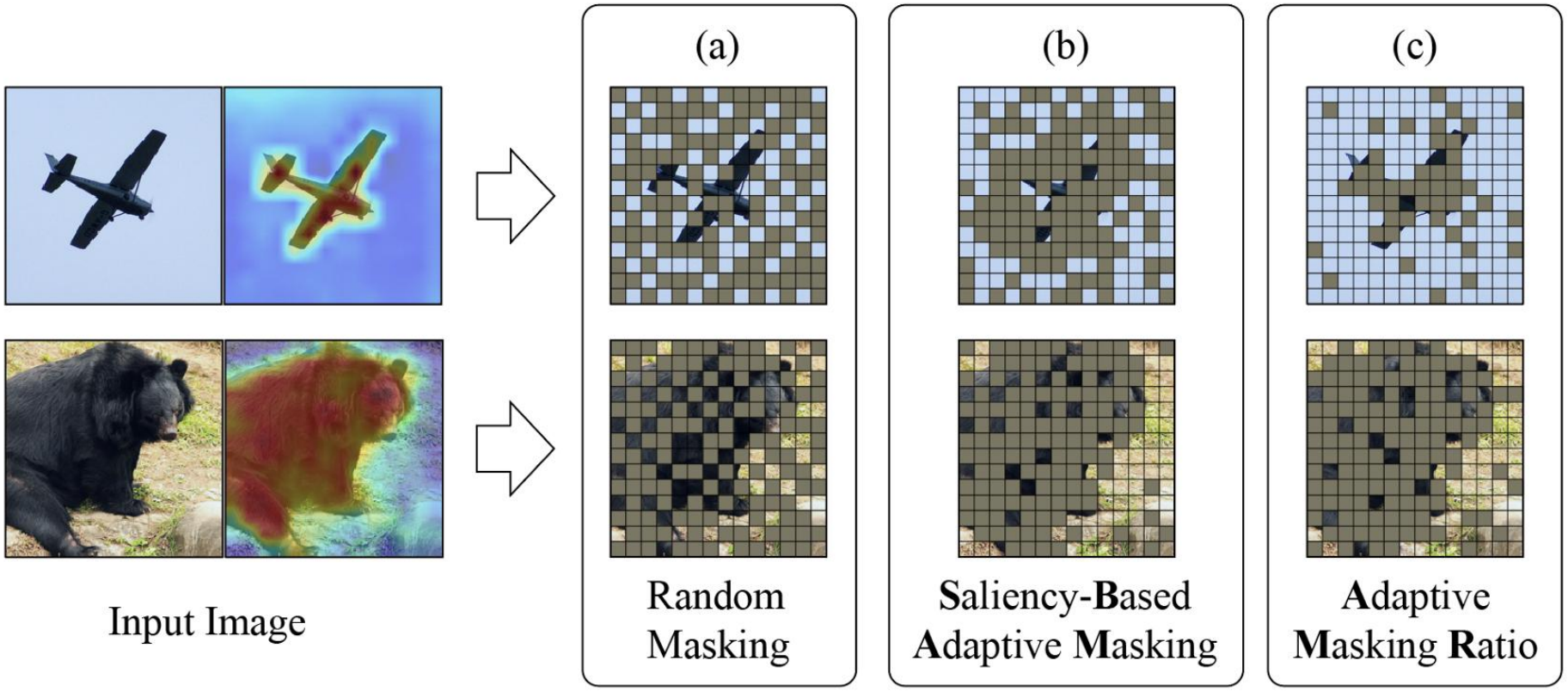
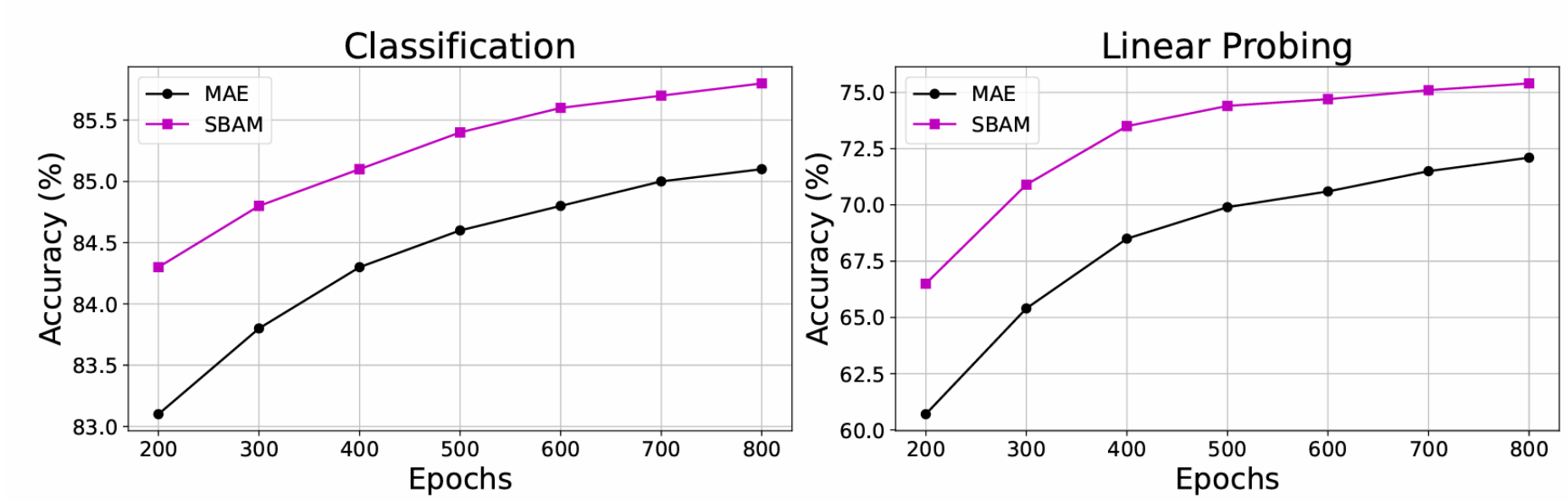


图 2. EMAE 方法示意图。整张图像首先被划分为 N 个图像块，然后并行策略将这 N 个图像块划分为 K 个不重叠的部分 $x_v^1,\ldots,x_v^K$，每一部分大小相同，每部分包含 N/K 个随机且不重叠的可见图像块。随后，每一部分被输入到编码器-解码器结构中执行 MIM 任务，生成 $x_m^1,\ldots,x_m^K$。此外，引入自一致性学习以将相同位置的重叠预测拉近一致。

AMR（自适应掩码比率策略）



AMR在预训练轮数中的性能比较。在ImageNet-1K数据集上，使用ViT-L预训练的图像分类精度的比较。左图展示了不同预训练轮数下的分类精度，右图展示了通过线性探测获得的精度。两项结果均表明，应用AMR显著提升了预训练性能，不仅在训练早期就取得了更高的精度，而且在收敛时仍保持领先。

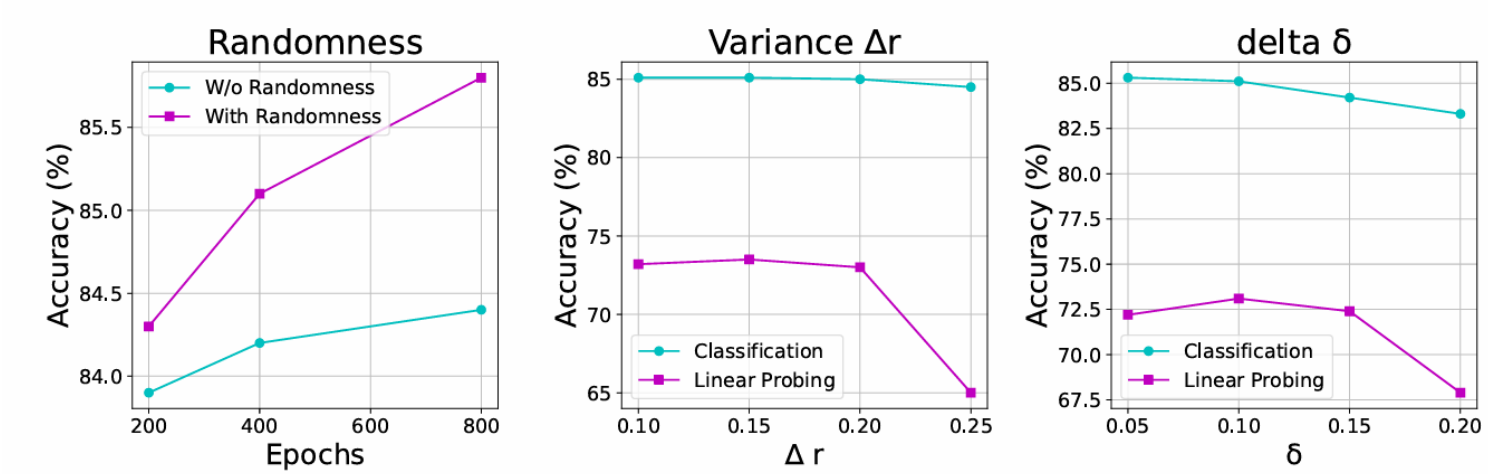
实验

Method	Baseline	Baseline+SBAM
MAE (ViT-L) [14]	84.3	85.1
MAE (ViT-B)	82.9	83.6
BootMAE (ViT-B) [11]	84.1	84.8
iBoT (ViT-B) [43]	71.5	74.4
CMAE (ViT-B) [17]	83.8	84.5

将SBAM应用于各种基线方法的综合性能结果。我们报告了在ImageNet-1K数据集上的图像分类微调精度的比较。SBAM在多种基线方法上的一致性能提升证明了其作为一种可扩展方法的有效性，能够增强各种MIM框架。

Method	Acc (%)
AM [25]	82.5
AMT [25]	82.8
SBAM	83.6

SBAM与最先进的掩码策略AMT 的比较评估



不同超参数对实验结果的影响

总结

本文提出了基于显著性自适应遮挡（SBAM）和自适应遮挡比例（AMR）的方法，为MIM领域带来了显著进展。本文的方法能够根据token的显著性动态调整遮挡位置与比例，从而提升了预训练的效率和在ImageNet-1K上的表现。同时，SBAM提出了一个全新的视角：将token之间的动态关系纳入遮挡策略，从而促使模型学习更关键的视觉表示。

不过我们也意识到，尽管SBAM结合了遮挡的随机性与显著性信息，它主要关注的是“高显著性token”的遮挡，可能会忽略图像中那些不显眼却具有上下文意义的区域。作者希望未来进一步平衡对显著token的关注与对低显著、但可能具有语义细节token的考虑，推动模型在理解整图语义方面更进一步。