JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

Masked Image Modeling with Local Multi-Scale Reconstruction

Haoqing Wang[1], Yehui Tang[1,2], Yunhe Wang[2], Jianyuan Guo[2], Zhi-Hong Deng[1], Kai Han[2]

[1]Peking University [2]Huawei Noah's Ark Lab

TUE-AM-203

LocalMIM: local multi-scale reconstruction

- For MIM models, thousands of GPU Hours for pre-training limit their industrial applications.

- Local reconstruction
  - we are the first to conduct reconstruction tasks at both lower and upper layers, which explicitly guide multiple layers to accelerate the representation learning.
- Multi-scale supervision
  - for both columnar and pyramidal architectures, the lower layers reconstruct the fine-scale supervision signals, and the upper layers reconstruct the coarse-scale ones.

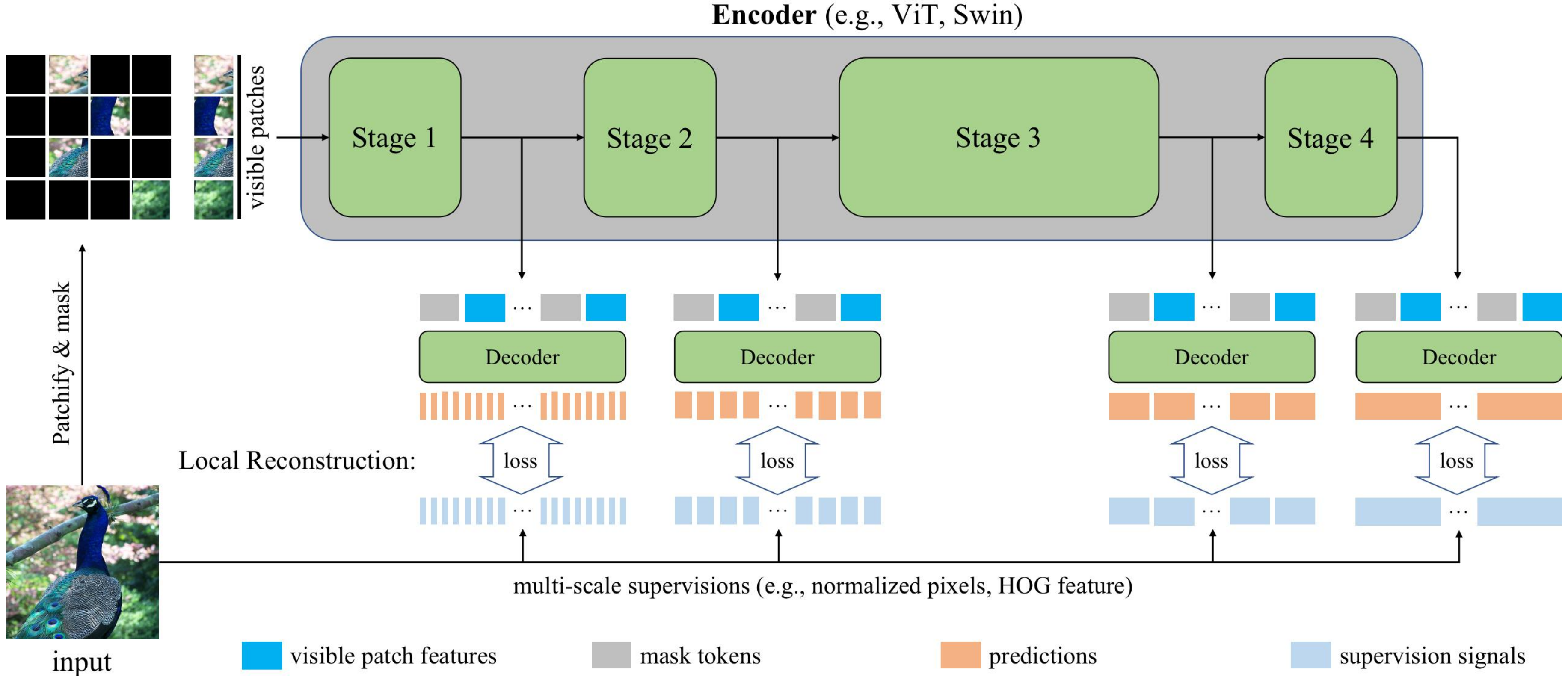
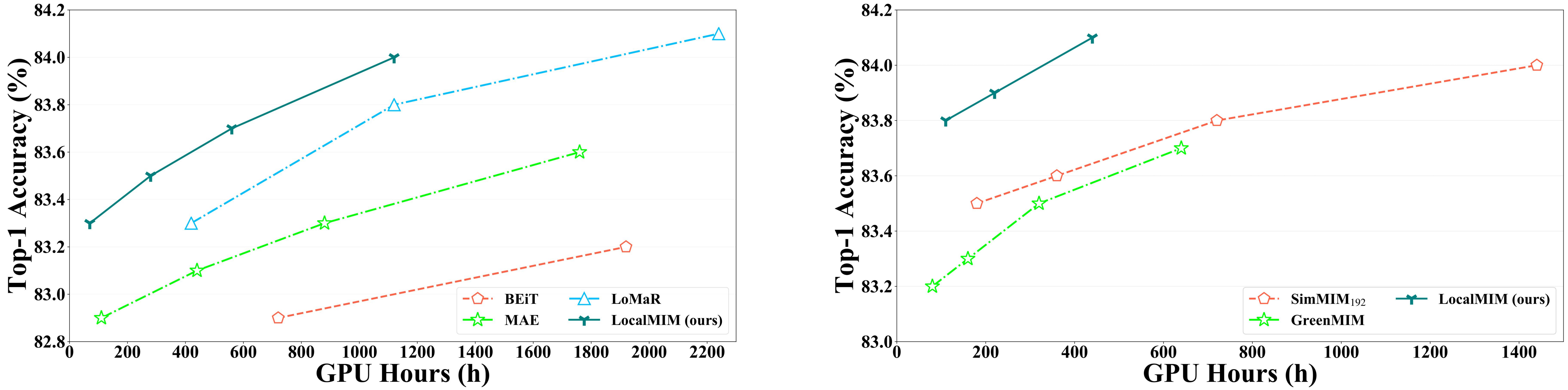


Figure: Overview of LocalMIM.

Performance

- LocalMIM is architecture-agnostic and can be used in both columnar and pyramidal architectures.
- On columnar ViT-B, LocalMIM achieves the best results of BEiT, MAE and MaskFeat with 27.4x, 3.1x and 5.6x acceleration respectively.
- On pyramidal Swin-B, LocalMIM achieves the best results of SimMIM_{192} and GreenMIM with 3.6x and 6.4x acceleration respectively.



(a) ViT-B (b) Swin-B

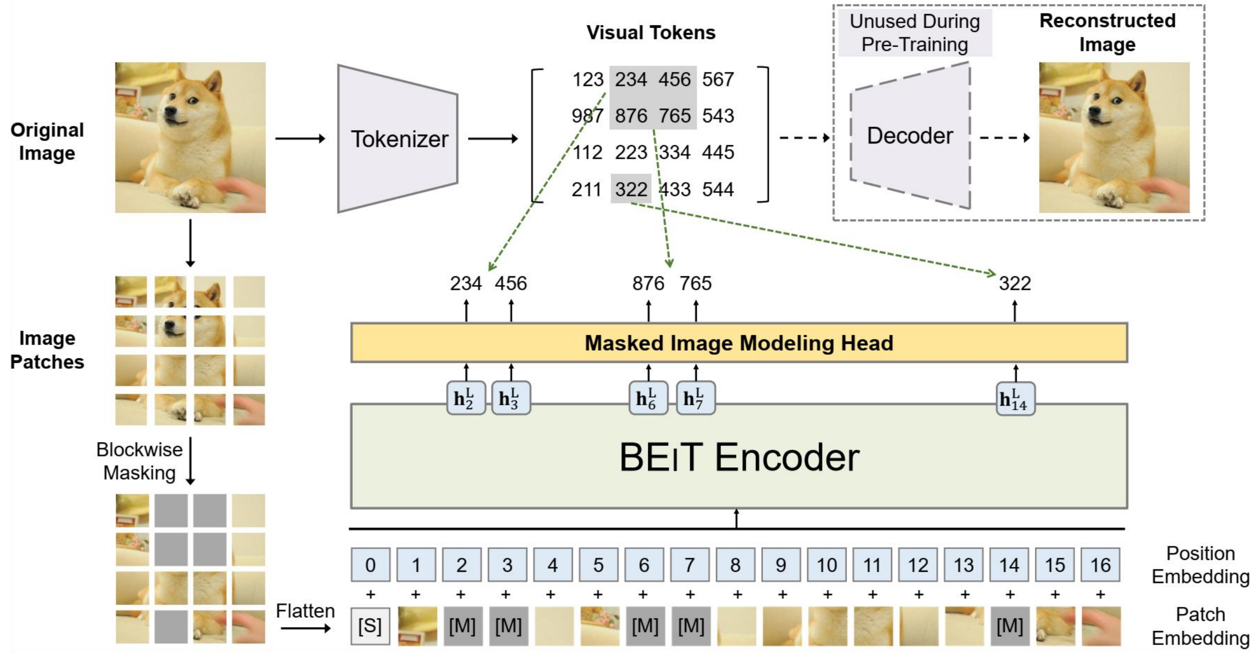
Figure: Top-1 fine-tuning accuracy on ImageNet-1K vs. Pre-training duration.

Background

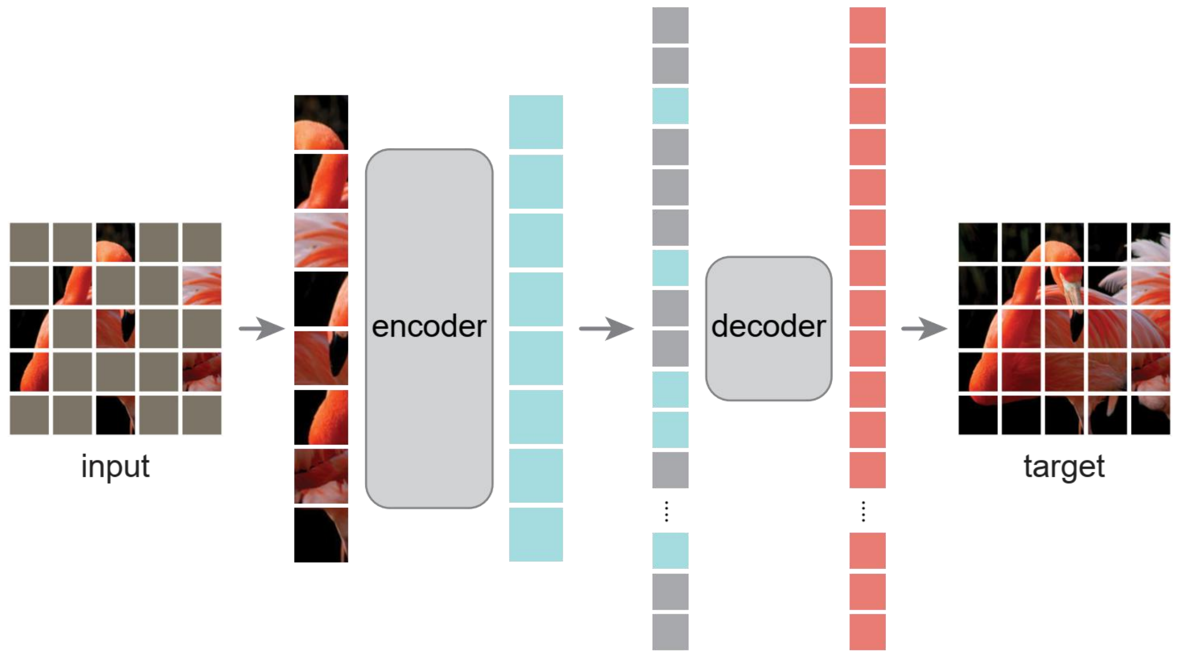
Masked Image Modeling: randomly mask some input parts and inference them based on other parts.

Classic works

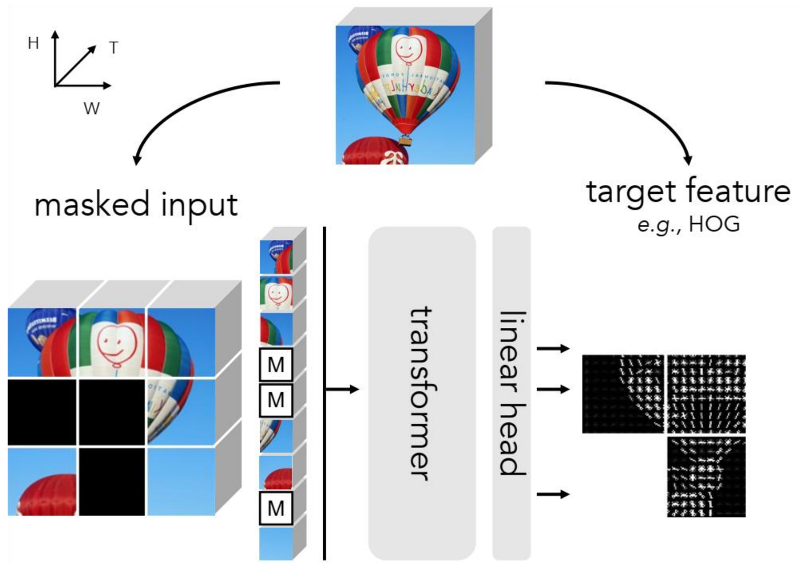
BEiT [Bao et al., 2022]



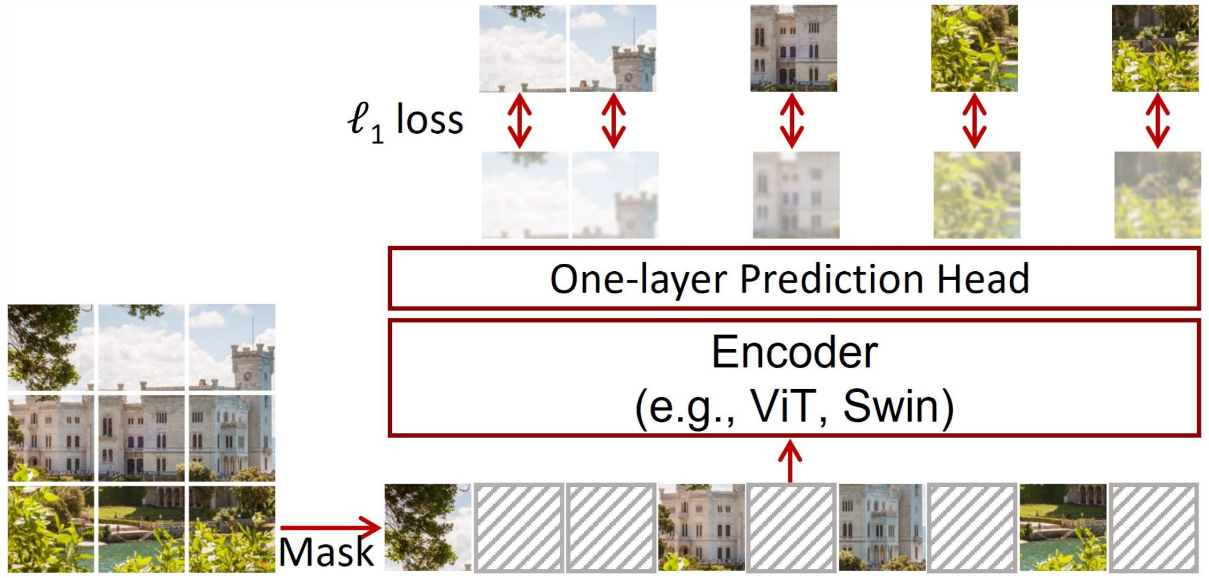
MAE [He et al., 2022]



MaskFeat [Wei et al., 2022]



SimMIM [Xie et al., 2022]



Background

Disadvantages: huge computational burden and slow pre-training process
The pre-training efficiency is an inevitable bottleneck limiting the industrial applications of MIM.

Existing works: accelerate the encoding process
1. the encoder only processes visible patches, e.g., MAE, GreenMIM.
2. shrinking the input resolution to lessen the input patches, e.g., LoMaR, UM-MAE, FastMIM.

None of them focus on the representation learning process itself!

Analysis

The lower layers of the encoder play the key role in the representation learning of MIM:
1) For pre-training, the well-learned lower layers can propagate knowledge to the upper ones and facilitate their learning.
2) For fine-tuning, the upper layers are typically tuned quickly to adapt the downstream task while the lower ones change more slowly and need to be well-learned during pre-training.

All existing MIM models only explicitly guide the top layer!

Analysis

Without explicit guidance, the inter-patch semantical relations on the lower layers can not be sufficiently learned.
It has the computational complexity with a quadratic dependence on patch number N, i.e., $\Theta(N^2)$.

Existing MIM models with global loss have small Normalized Mutual Information (NMI) at lower layers, which means their patches have less query-adaptive attentions.

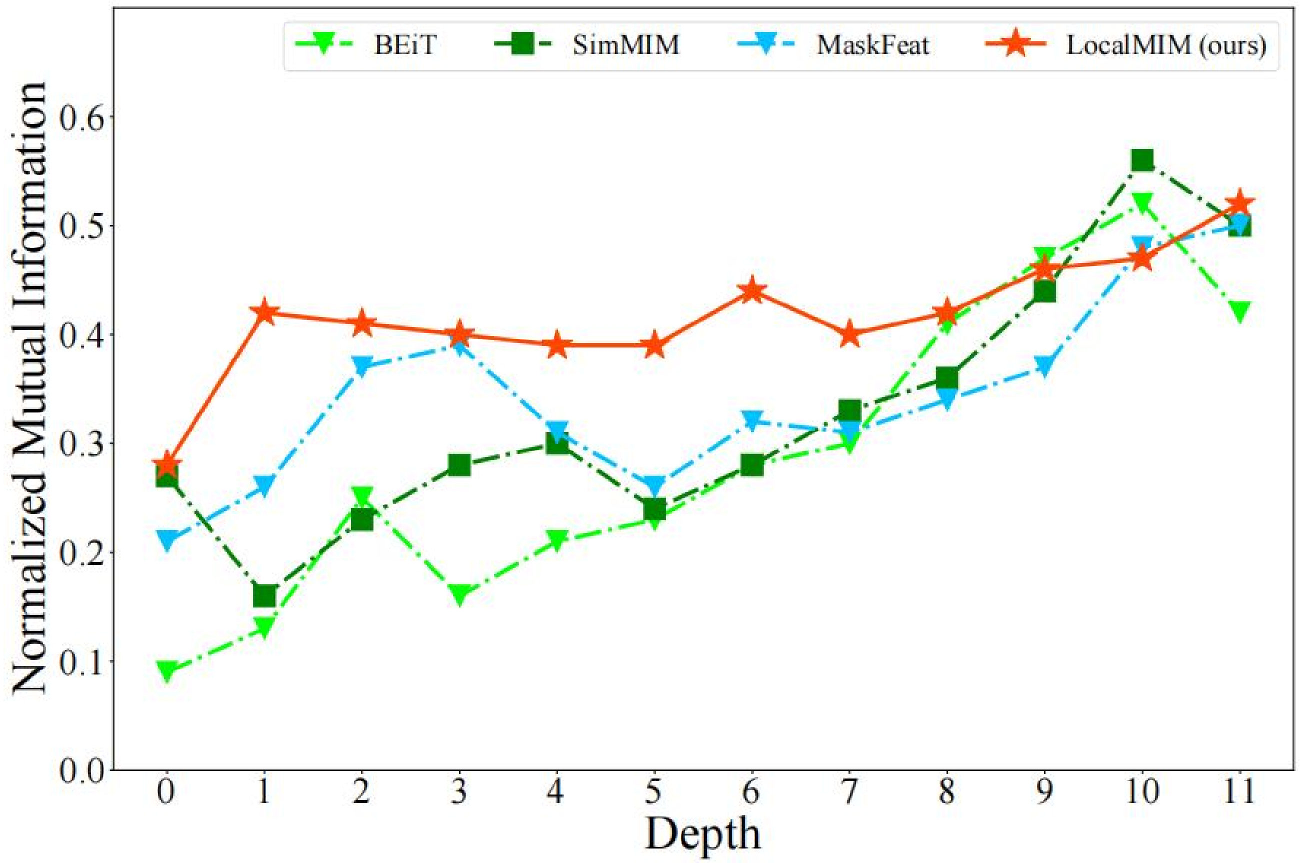


Figure: NMI between query and key patches at each layer

Model

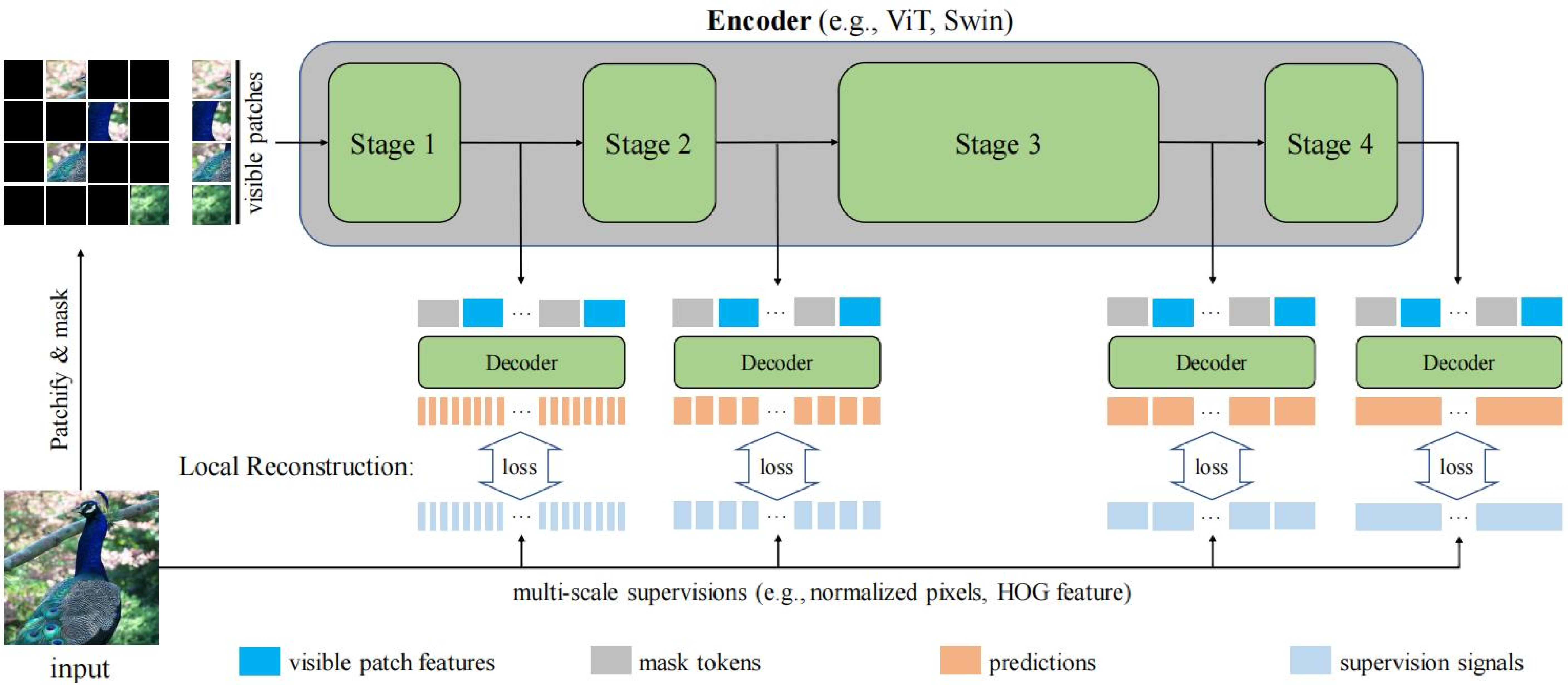


Figure: Overview of LocalMIM.

1. we are the *first* to conduct reconstruction tasks at multiple layers.
2. we are the *first* to use multiple scale supervision signals, where the lower layers reconstruct the fine-scale supervisions and the upper layers reconstruct the coarse-scale ones.

Model

Input:

$$x \in \mathbb{R}^{H \times W \times C}$$

Supervisions:

$$y_i = \pi(x_i)$$

where $\{x_i \in \mathbb{R}^{p \times p \times C}\}_{i=1}^{HW/p^2}$ are non-overlapping patches with the scale of $\frac{H}{p} \times \frac{W}{p}$, π is the feature descriptor, e,g., codebook, HOG, pixel normalization.

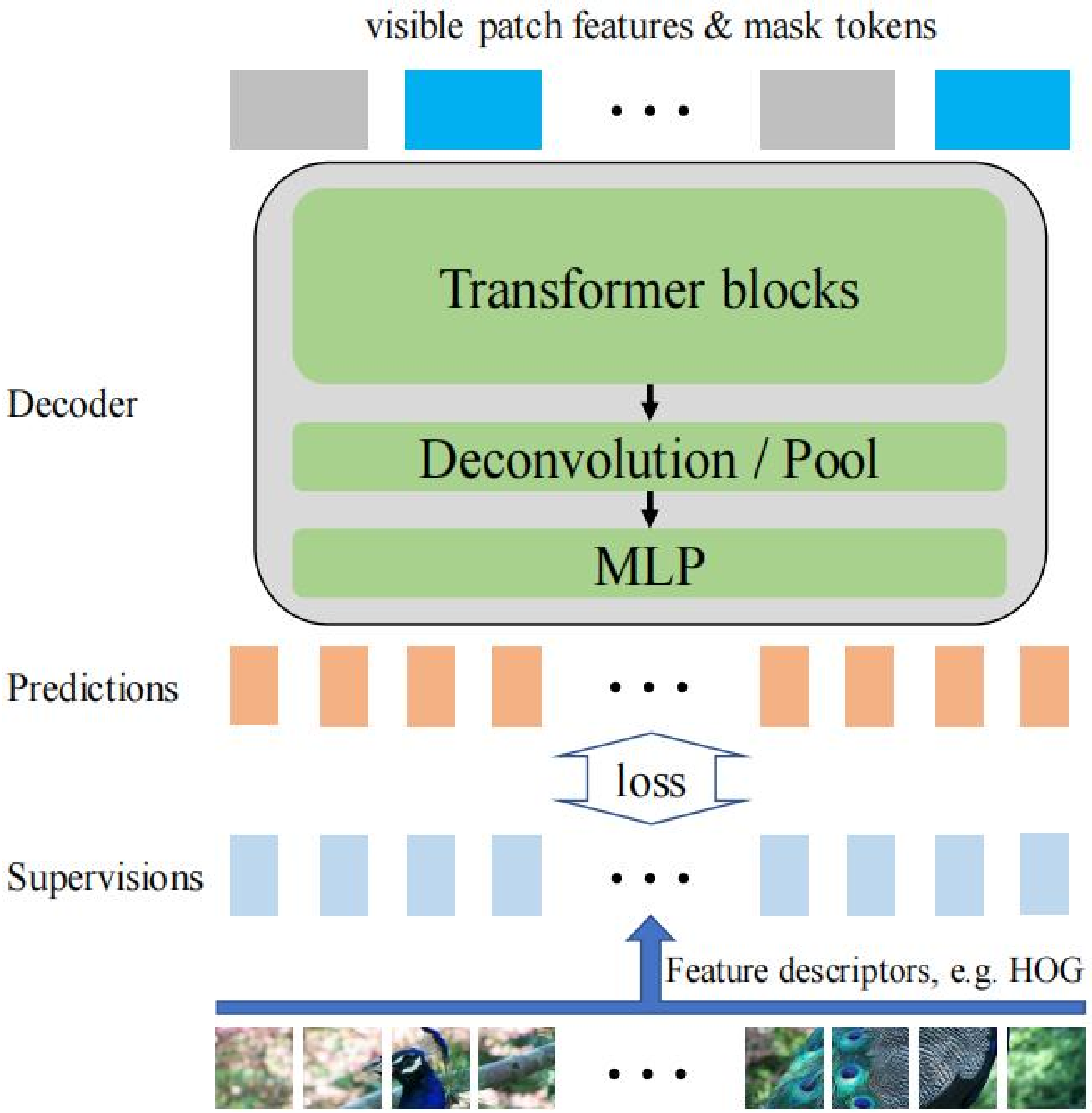


Figure: reconstruction process under a scale.

Decoder: Transformer block + rescale + MLP

The tiny decoders have only one Transformer block with small embedding dim and few attention heads.

Experiments

1. Classification on ImageNet-1K

ViT-B:

BEiT 27.4x

MAE 3.1x

MaskFeat 5.6x

Swin-B:

SimMIM 3.6x

GreenMIM 6.4x

Model	**Backbone**	**# Params**	**PT Epoch**	**GPU Hours/Ep.**	**Total GPU Hours**	**Acc**
Scratch, ViT	ViT-B	86M	0	1.5	-	82.3
Scratch, Swin	Swin-B	88M	0	2.4	-	83.5
MoCo v3 [12]	ViT-B	86M	600	-	-	83.2
DINO [7]	ViT-B	86M	300	-	-	82.8
BEiT [2]	ViT-B	86M	800	2.4	1920	83.2
iBOT [67]	ViT-B	86M	400	10.1	4040	83.8
MAE [24]	ViT-B	86M	800	1.1	880	83.3
MAE [24]	ViT-B	86M	1600	1.1	1760	83.6
MAE [24]	ViT-L	307M	1600	1.7	2720	85.9
MaskFeat [57]	ViT-B	86M	1600	3.9	6240	84.0
CAE [11]	ViT-B	86M	800	2.8	2240	83.6
LoMaR† [8]	ViT-B	86M	1600	1.4	2240	84.1
data2Vec† [1]	ViT-B	86M	800	3.0	2400	84.2
PeCo [17]	ViT-B	86M	800	-	-	84.5
LocalMIM-HOG	ViT-B	86M	100	0.7	70	83.3
LocalMIM-HOG	ViT-B	86M	1600	0.7	1120	84.0
LocalMIM-HOG	ViT-L	307M	800	1.0	800	85.8
SimMIM_{192} [60]	Swin-B	88M	800	1.8	1440	84.0
SimMIM_{192} [60]	Swin-L	197M	800	3.0	2400	85.4
GreenMIM [31]	Swin-B	88M	800	0.8	640	83.7
GreenMIM [31]	Swin-L	197M	800	1.4	1120	85.1
LocalMIM-Pixel	Swin-B	88M	100	1.0	100	83.7
LocalMIM-HOG	Swin-B	88M	100	1.1	110	83.8
LocalMIM-Pixel	Swin-B	88M	400	1.0	400	84.0
LocalMIM-HOG	Swin-B	88M	400	1.1	440	84.1
LocalMIM-HOG	Swin-L	197M	800	1.6	1280	85.6

Figure: Top-1 fine-tuning accuracy on ImageNet-1K.

Experiments

2. Segmentation and Detection

Model	PT Epoch	PT Hours	mIoU
Supervised	-	-	47.4
MoCo v3 [12]	300	-	47.3
BEiT [2]	800	1920	47.1
MAE [24]	1600	1760	48.1
MaskFeat [57]	1600	6240	48.8
PeCo [17]	800	-	48.5
CAE [11]	800	2240	48.8
LocalMIM-HOG	1600	1120	49.5

Figure: Semantic segmentation on ADE20K

Model	PT Epoch	PT Hours	AP^b	AP^m
Supervised	300	840	48.5	43.2
$SimMIM_{192}$ [60]	800	1440	50.4	44.4
GreenMIM [31]	800	640	50.0	44.1
LocalMIM-HOG	400	440	50.7	44.9

Figure: Object detection and instance segmentation on COCO

1. LocalMIM significantly outperforms supervised pre-training.
2. LocalMIM achieves better performance than other MIM models with less pre-training burden.

Experiments

3. Visualization of the attention maps

1) For object-centric images, LocalMIM can distinguish the foreground object from the background.

2) For multi-object images, LocalMIM can effectively separate different objects without any task-specific supervision, which means the attention maps are query-adaptive.

3) The patches at lower layers typically more focus on their neighboring regions, while those at upper layers attend to a wide range of semantically related regions.

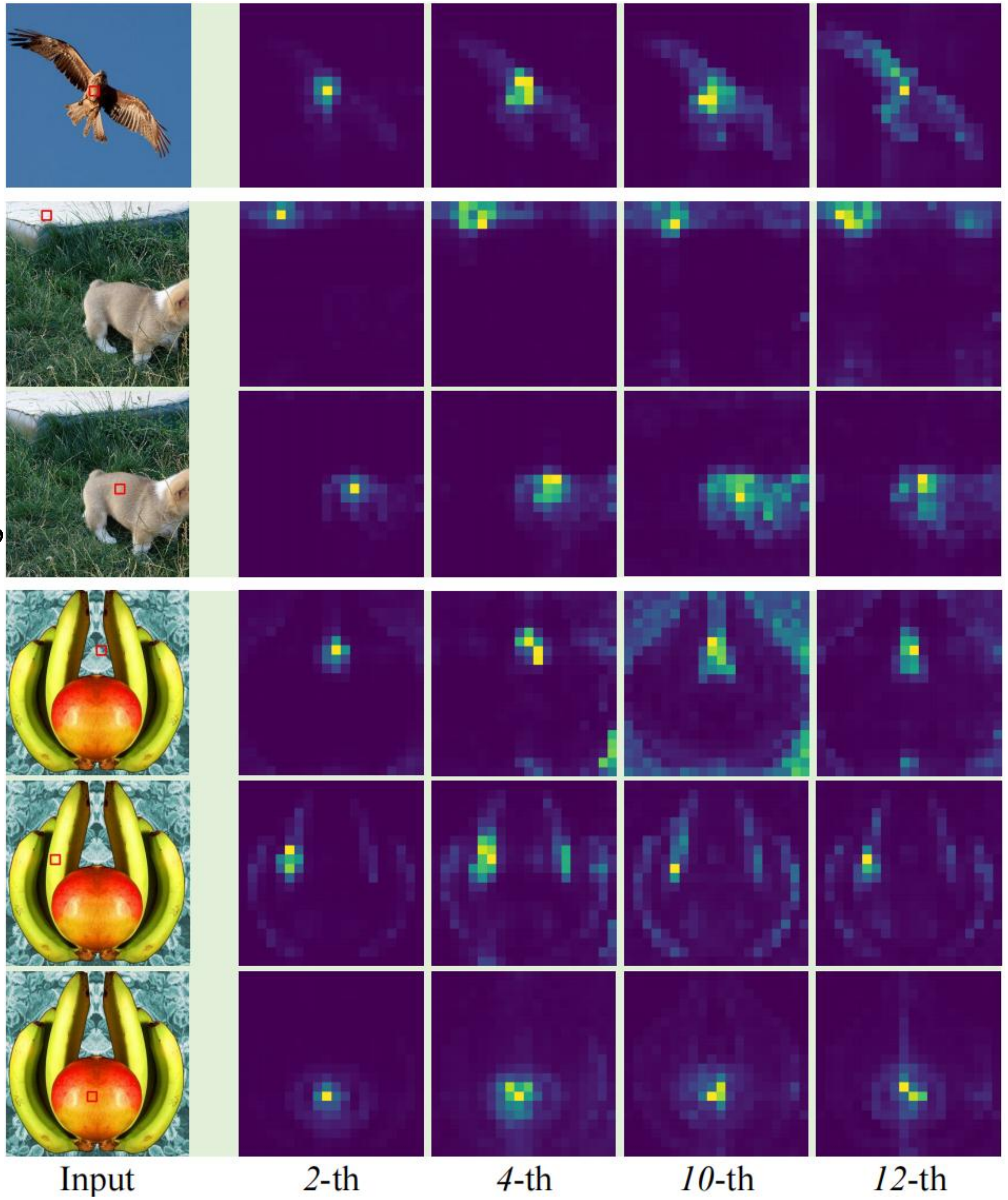


Figure: Visualization of the attention maps for different query points, marked with red boxes.

Experiments

4. Gradient-isolated pre-training

model	backbone	GPU Hours/Ep.	acc
LocalMIM	ViT-B	0.7	83.3
w/ isolated grad		0.7	83.0
LocalMIM	Swin-B	1.1	83.8
w/ isolated grad		1.1	83.7

Surprisingly, the gradient-isolated training achieves similar performance to global back-propagation.

Thank you!

Paper:https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_Masked_Image_Modeling_With_Local_Multi-Scale_Reconstruction_CVPR_2023_paper.pdf

Code: https://github.com/huawei-noah/Efficient-Computing/tree/master/Self-supervised/LocalMIM