What to Hide from Your Students: Attention-Guided Masked Image Modeling

Authors: I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, N. Komodakis

Narrator: Ioannis Kakogeorgiou

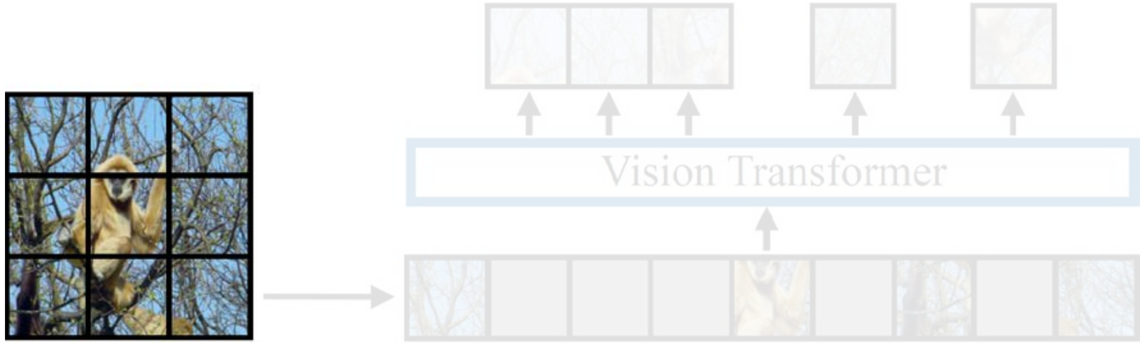
National Technical University of Athens
ATHENA Research & Innovation Information Technologies
University of Crete
IARAI institute of advanced research in artificial intelligence
valeo.ai
FORTH

Code: https://github.com/gkakogeorgiou/attmask

ECCV TEL AVIV 2022

Scope: Self-supervised learning of Vision Transformers (ViT) via Masked Image Modeling (MIM)

- Divide an input image into patches tokens
- Mask a portion of the input patch tokens
- Train a Vision Transformer to reconstruct them

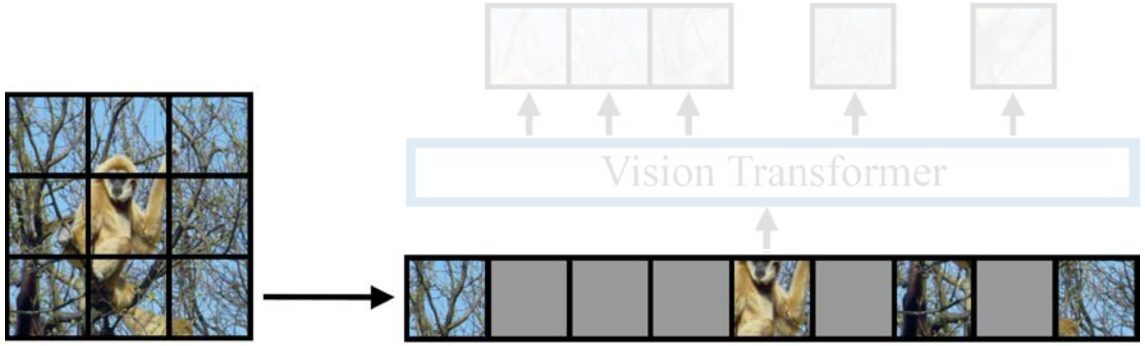


Bao et al. BEiT: BERT Pre-Training of Image Transformers ICLR, 2022.
Xie et al. SimMIM: A Simple Framework for Masked Image Modeling CVPR, 2022.

Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022.
He et al. Masked Autoencoders Are Scalable Vision Learners CVPR, 2022.

Scope: Self-supervised learning of Vision Transformers (ViT) via Masked Image Modeling (MIM)

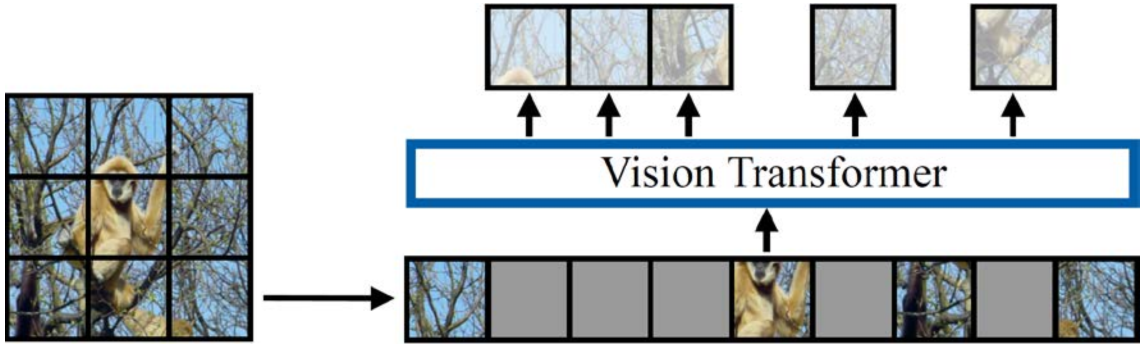
- Divide an input image into patches tokens
- Mask a portion of the input patch tokens
- Train a Vision Transformer to reconstruct them



Bao et al. BEiT: BERT Pre-Training of Image Transformers ICLR, 2022.
Xie et al. SimMIM: A Simple Framework for Masked Image Modeling CVPR, 2022.
Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022.
He et al. Masked Autoencoders Are Scalable Vision Learners CVPR, 2022.

Scope: Self-supervised learning of Vision Transformers (ViT) via Masked Image Modeling (MIM)

- Divide an input image into patches tokens
- Mask a portion of the input patch tokens
- Train a Vision Transformer to reconstruct them



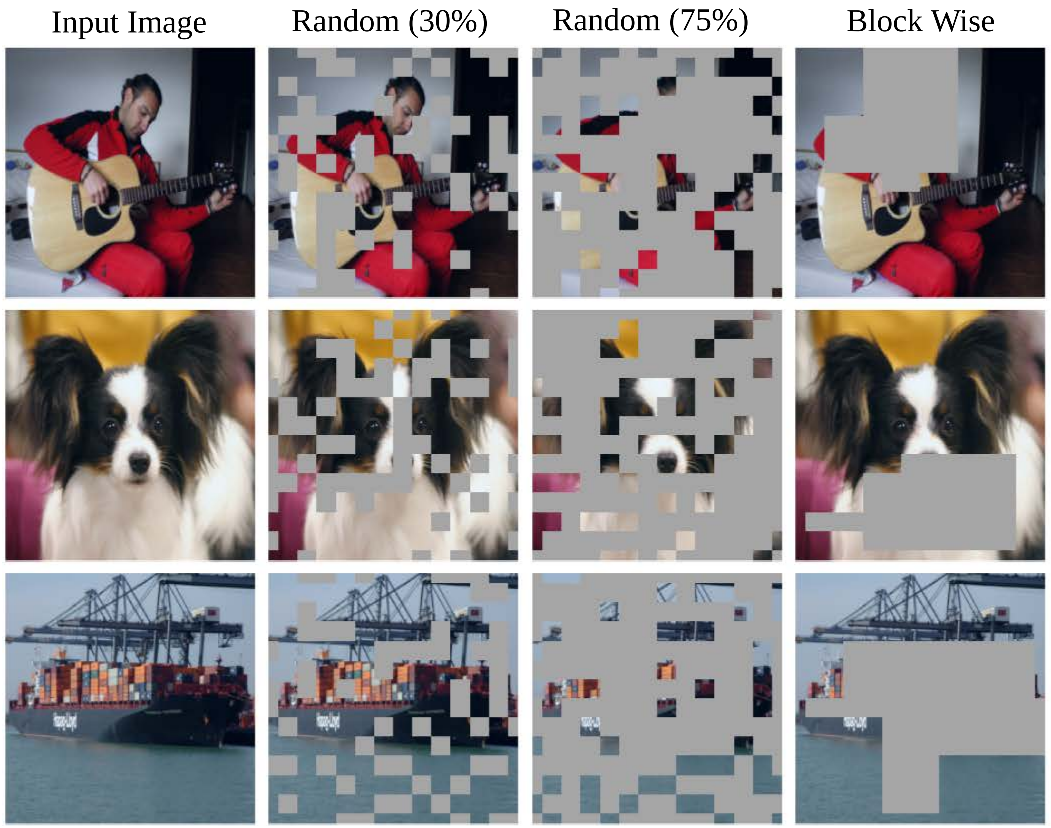
Bao et al. BEiT: BERT Pre-Training of Image Transformers ICLR, 2022.
Xie et al. SimMIM: A Simple Framework for Masked Image Modeling CVPR, 2022.
Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022.
He et al. Masked Autoencoders Are Scalable Vision Learners CVPR, 2022.

Focus: Which patch tokens to mask?

- **Not well explored;** prior works use **(block-wise) random** token masking

Focus: Which patch tokens to mask?

- **Not well explored;** prior works use **(block-wise) random** token masking
  - Less likely to hide "interesting" parts → **easy reconstruction**
  - Compensating with extreme masking (e.g., 75% of tokens) → **overly aggressive**



He et al. Masked Autoencoders Are Scalable Vision Learners CVPR, 2022.

Bao et al. BEiT: BERT Pre-Training of Image Transformers ICLR, 2022.

Our Approach: Attention-guided token masking (AttMask)

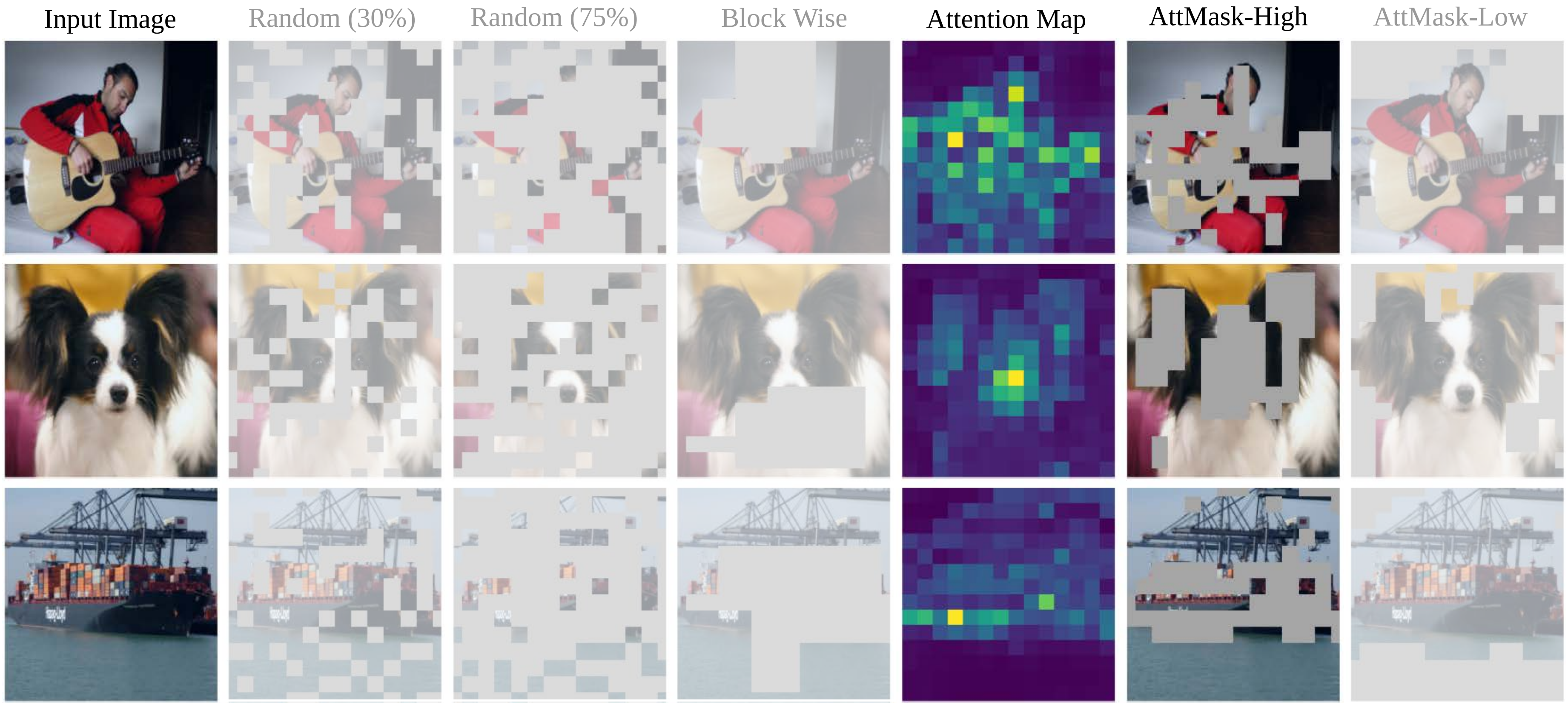
- Leverage ViT's self-attention to mask tokens

Our Approach: Attention-guided token masking (AttMask)

- Leverage ViT's self-attention to mask tokens
  - × **AttMask-Low**: masks low-attended tokens (essentially background)
    → **very easy** reconstruction task → **degrades performance**

Our Approach: Attention-guided token masking (AttMask)

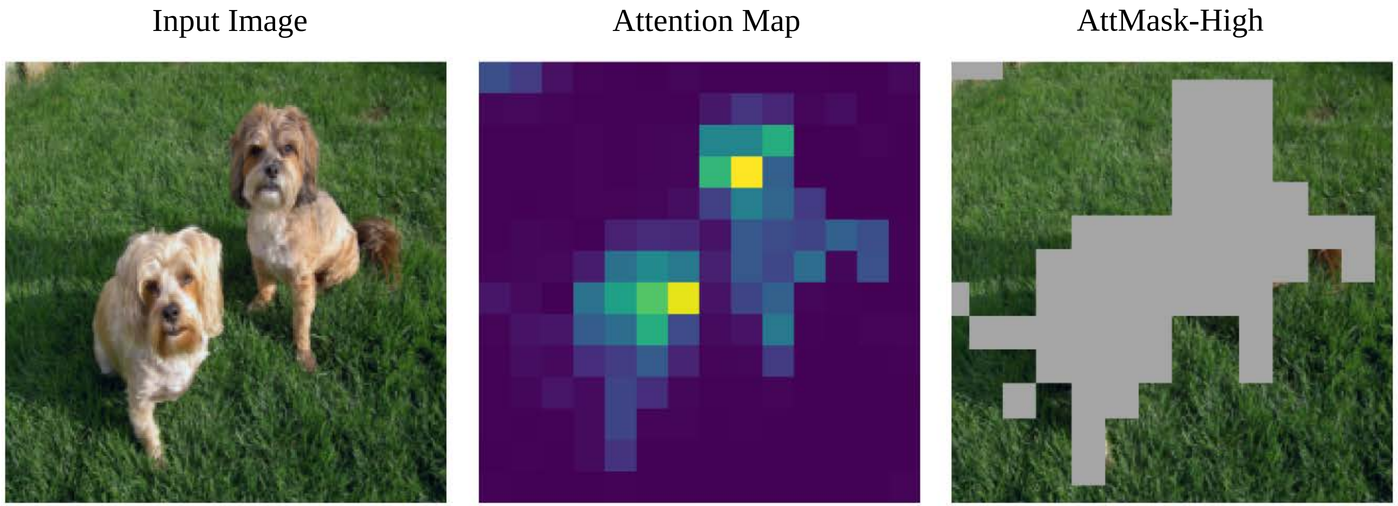
- Leverage ViT's self-attention to mask tokens
  - ✓ **AttMask-High**: masks highly-attended tokens (essentially foreground)
    → **very challenging** reconstruction task → **boosts performance**



Our Approach: Attention-guided token masking (AttMask)

- Leverage ViT's self-attention to mask tokens
  - ✓ **AttMask-High**: masks highly-attended tokens (essentially foreground) → **very challenging** reconstruction task → **boosts performance**

Perhaps overly aggressive for high mask ratios

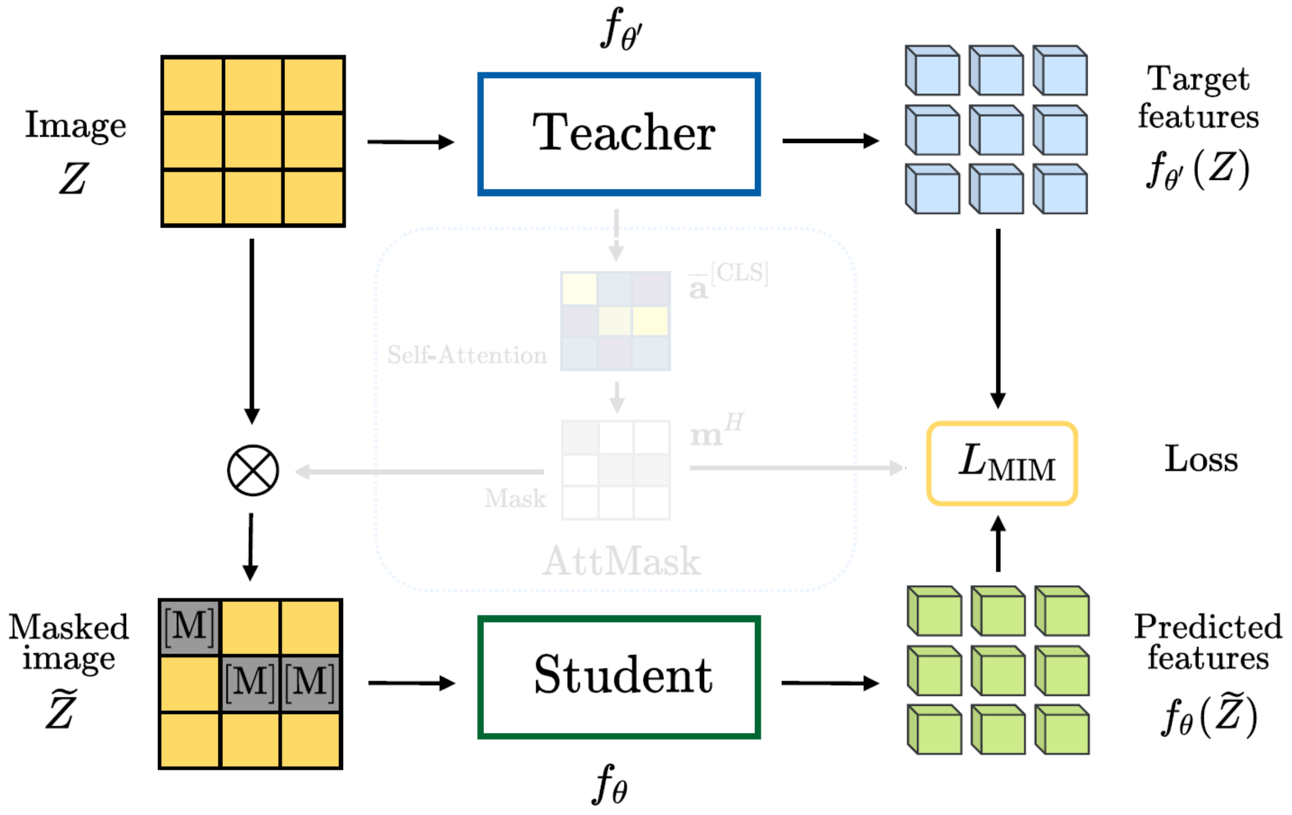


Our Approach: Attention-guided token masking (AttMask)

- Leverage ViT's self-attention to mask tokens
  - ✓ **AttMask-High**: masks highly-attended tokens (essentially foreground)
    → **very challenging** reconstruction task → **boosts performance**
  - ✓ **AttMask-Hint**: masks highly-attended tokens but leaves some hints
    → **provides hints** for the identity of the masked object → **boosts performance**

Input Image | Attention Map | AttMask-High | AttMask-Hint

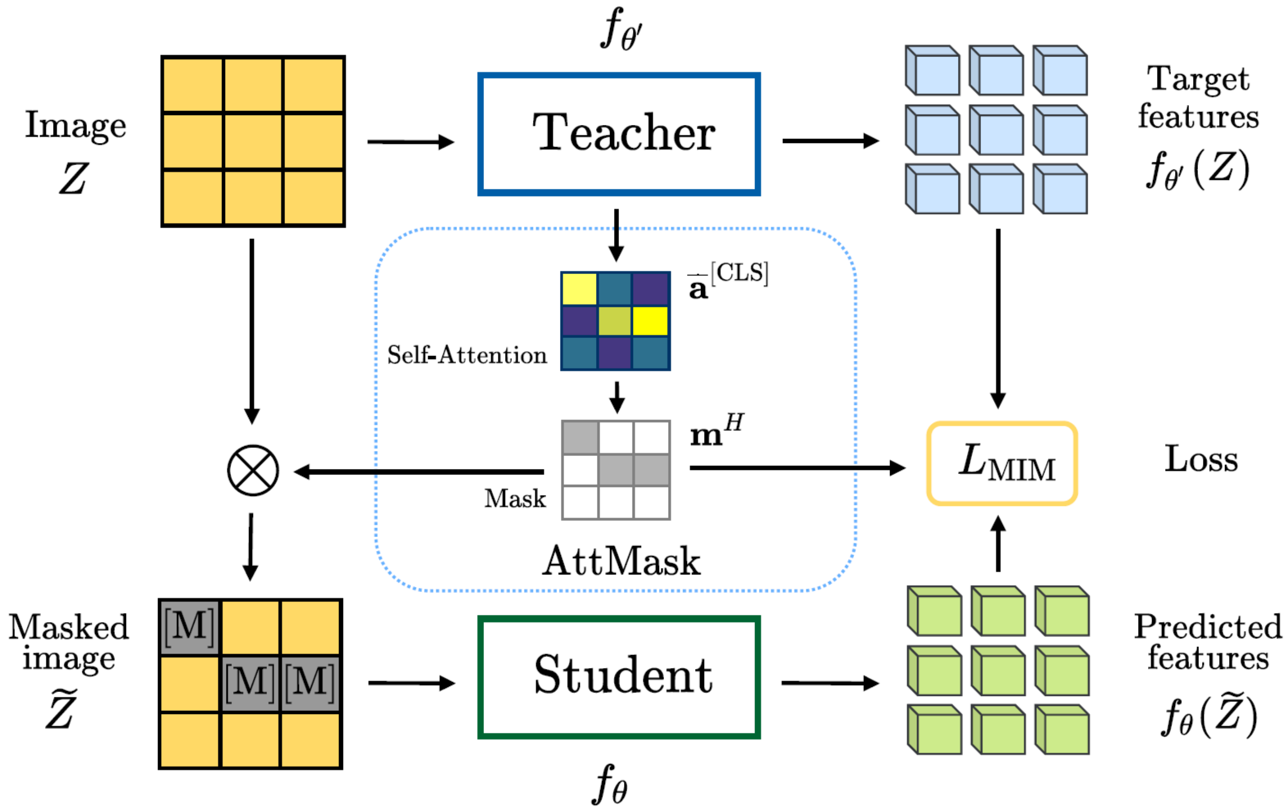
Incorporating AttMask into distillation-based methods



- We exhibit AttMask in the context of distillation-based MIM, such as iBOT [1]
- The teacher transformer encoder sees the entire image and generates the attention map
- The student sees only the masked image and solves the reconstruction task
- AttMask thus incurs zero additional cost

[1] Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022

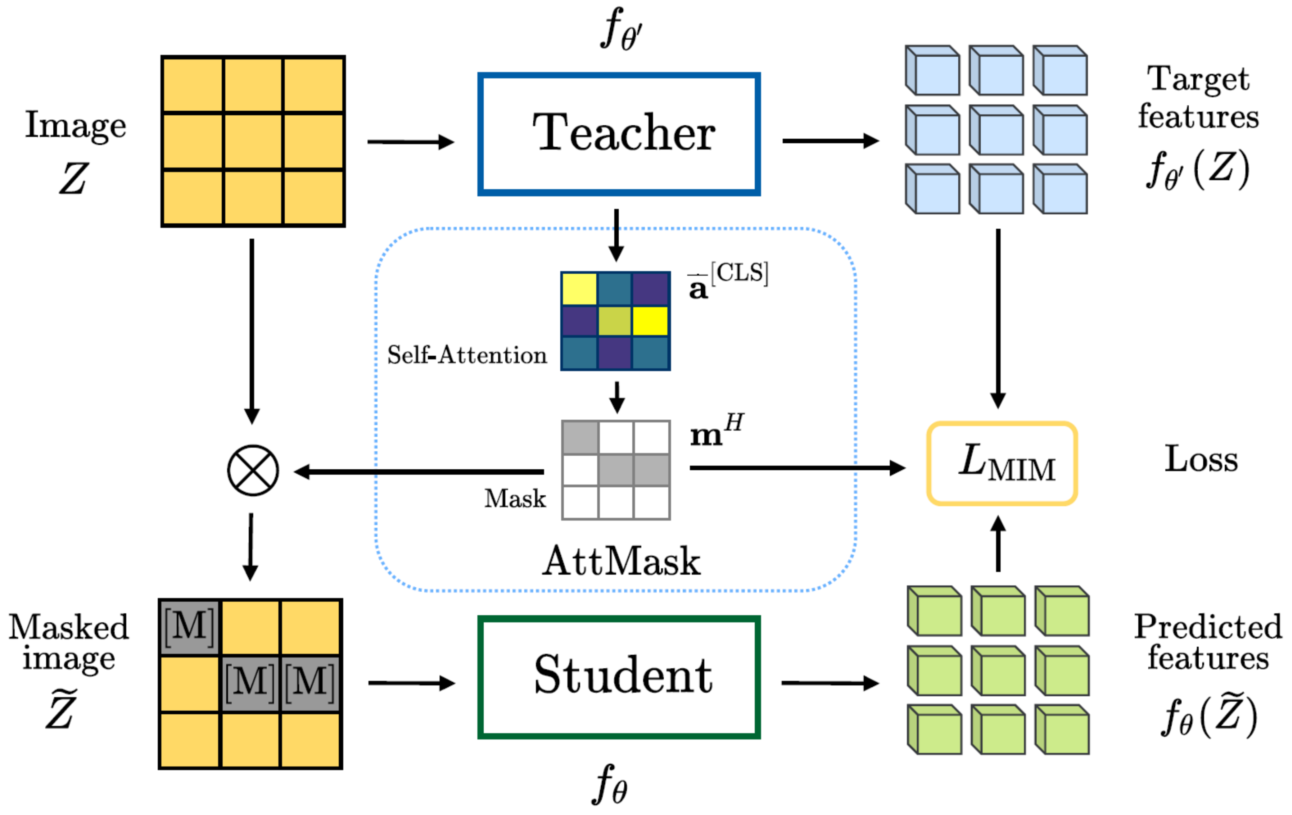
Incorporating AttMask into distillation-based methods



- We exhibit AttMask in the context of distillation-based MIM, such as iBOT [1]
- The teacher transformer encoder sees the entire image and generates the attention map
- The student sees only the masked image and solves the reconstruction task
- AttMask thus incurs zero additional cost

[1] Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022

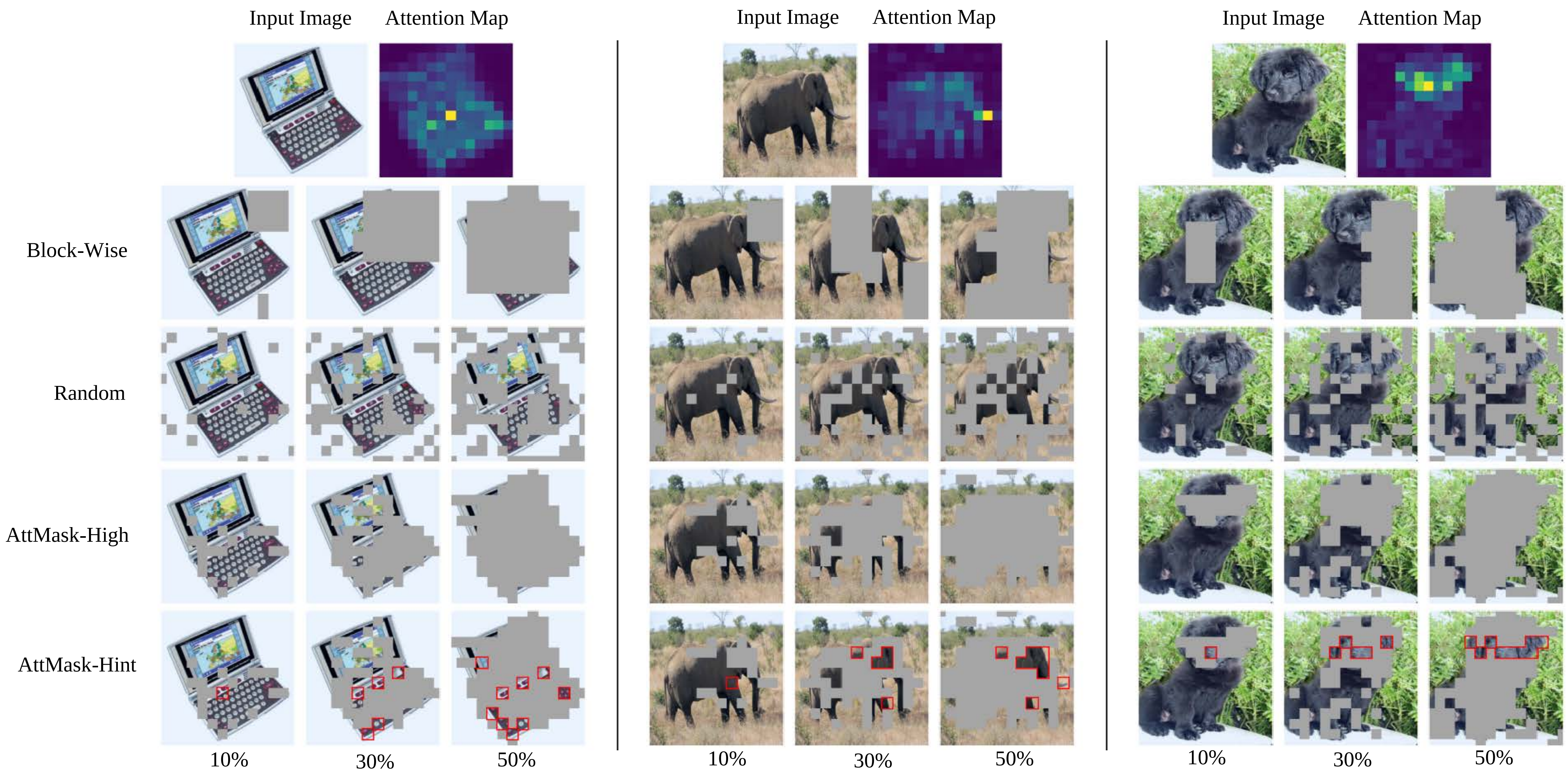
Incorporating AttMask into distillation-based methods



- We exhibit AttMask in the context of distillation-based MIM, such as iBOT [1]
- The teacher transformer encoder sees the entire image and generates the attention map
- The student sees only the masked image and solves the reconstruction task
- AttMask thus incurs zero additional cost

[1] Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022

Qualitative examination of masking strategies



Evaluating token masking strategies by pre-training on 20% of ImageNet-1k

Top-1 accuracy for k-NN and linear probing on ImageNet validation set as well as fine-tuning on CIFAR10/100.

iBOT Masking	Ratio (%)	ImageNet-1k		CIFAR10	CIFAR100
		k-NN	Linear	Fine-tuning	
Random Block-Wise†	10-50	46.7	56.4	98.0	86.0
Random‡	75	47.3	55.5	97.7	85.5
Random	10-50	47.8	56.7	98.0	86.1
AttMask-Low (ours) ✗	10-50	44.0	53.4	97.6	84.6
AttMask-Hint (ours) ✓	10-50	49.5	57.5	98.1	**86.6**
AttMask-High (ours) ✓	10-50	**49.7**	**57.9**	**98.2**	**86.6**

†: default iBOT masking strategy from BEiT ‡: aggressive random masking strategy from MAE

✓ AttMask-High improves iBOT by +3% on k-NN and +1.5% on linear probing

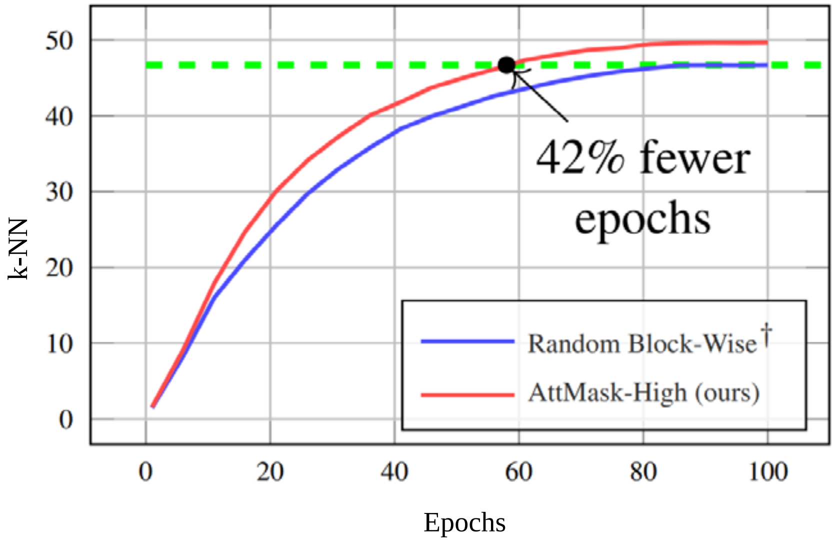
✓ AttMask-High accelerates the learning process

Evaluating token masking strategies by pre-training on 20% of ImageNet-1k

Top-1 accuracy for k-NN and linear probing on ImageNet validation set as well as fine-tuning on CIFAR10/100.

iBOT Masking	Ratio (%)	ImageNet-1k		CIFAR10	CIFAR100
		k-NN	Linear	Fine-tuning	
Random Block-Wise†	10-50	46.7	56.4	98.0	86.0
Random‡	75	47.3	55.5	97.7	85.5
Random	10-50	47.8	56.7	98.0	86.1
AttMask-Low (ours) ×	10-50	44.0	53.4	97.6	84.6
AttMask-Hint (ours) ✓	10-50	49.5	57.5	98.1	**86.6**
AttMask-High (ours) ✓	10-50	**49.7**	**57.9**	**98.2**	**86.6**

†: default iBOT masking strategy from BEiT ‡: aggressive random masking strategy from MAE



- ✓ AttMask-High improves iBOT by +3% on k-NN and +1.5% on linear probing
- ✓ AttMask-High accelerates the learning process

Evaluating by pre-training on different percentage (%) of ImageNet-1k for 100 epochs

Top-1 k-NN accuracy on ImageNet-1k validation for iBOT pre-training on different percentage (%) of ImageNet-1k

% ImageNet-1k	5	10	20	100
Random Block-Wise†	15.7	31.9	46.7	71.5
AttMask-High (ours)	**17.5**	**33.8**	**49.7**	72.5

†: default iBOT masking strategy from BEiT

Top-1 accuracy on ImageNet validation set. (a) k-NN and linear probing using the full ImageNet training set; (b) k-NN using only ν ∈ {1, 5, 10, 20} examples per class. Pre-training on 100% ImageNet-1k for 100 epochs.

Method	Full		Few Examples			
	k-NN	Linear	$\nu = 1$	5	10	20
DINO	70.9	74.6				
MST	72.1	75.0				
iBOT	71.5	74.4	32.9	47.6	52.5	56.4
iBOT+AttMask-High	72.5	75.7	37.1	51.3	55.7	59.1
iBOT+AttMask-Hint	**72.8**	**76.1**	**37.6**	**52.2**	**56.4**	**59.6**

Improved performance when:

- ✓ Pretraining with fewer data
- ✓ Pretraining on the full ImageNet-1k (+1.3% on k-NN and +1.5% on linear probing)
- ✓ Evaluating using only 1, 5, 10 or 20 samples per class for the k-NN classifier (more than +3% on low shot k-NN)

Evaluating by pre-training on different percentage (%) of ImageNet-1k for 100 epochs

Top-1 k-NN accuracy on ImageNet-1k validation for iBOT pre-training on different percentage (%) of ImageNet-1k

% ImageNet-1k	5	10	20	100
Random Block-Wise†	15.7	31.9	46.7	71.5
AttMask-High (ours)	**17.5**	**33.8**	**49.7**	**72.5**

†: default iBOT masking strategy from BEiT

Top-1 accuracy on ImageNet validation set. (a) k-NN and linear probing using the full ImageNet training set; (b) k-NN using only ν ∈ {1, 5, 10, 20} examples per class. Pre-training on 100% ImageNet-1k for 100 epochs.

Method	Full		Few Examples			
	k-NN	Linear	$\nu = 1$	5	10	20
DINO	70.9	74.6				
MST	72.1	75.0				
iBOT	71.5	74.4	32.9	47.6	52.5	56.4
iBOT+AttMask-High	72.5	75.7	37.1	51.3	55.7	59.1
iBOT+AttMask-Hint	**72.8**	**76.1**	**37.6**	**52.2**	**56.4**	**59.6**

Improved performance when:

- ✓ Pretraining with fewer data
- ✓ Pretraining on the full ImageNet-1k (+1.3% on k-NN and +1.5% on linear probing)
- ✓ Evaluating using only 1, 5, 10 or 20 samples per class for the k-NN classifier (more than +3% on low shot k-NN)

Evaluating by pre-training on different percentage (%) of ImageNet-1k for 100 epochs

Top-1 k-NN accuracy on ImageNet-1k validation for iBOT pre-training on different percentage (%) of ImageNet-1k

% ImageNet-1k	5	10	20	100
Random Block-Wise†	15.7	31.9	46.7	71.5
AttMask-High (ours)	**17.5**	**33.8**	**49.7**	**72.5**

†: default iBOT masking strategy from BEiT

Top-1 accuracy on ImageNet validation set. (a) k-NN and linear probing using the full ImageNet training set; (b) k-NN using only $\nu \in \{1, 5, 10, 20\}$ examples per class. Pre-training on 100% ImageNet-1k for 100 epochs.

Method	Full		Few Examples			
	k-NN	Linear	$\nu = 1$	5	10	20
DINO	70.9	74.6				
MST	72.1	75.0				
iBOT	71.5	74.4	32.9	47.6	52.5	56.4
iBOT+AttMask-High	72.5	75.7	37.1	51.3	55.7	59.1
iBOT+AttMask-Hint	**72.8**	**76.1**	**37.6**	**52.2**	**56.4**	**59.6**

Improved performance when:

- ✓ Pretraining with fewer data
- ✓ Pretraining on the full ImageNet-1k (+1.3% on k-NN and +1.5% on linear probing)
- ✓ Evaluating using only 1, 5, 10 or 20 samples per class for the k-NN classifier (more than +3% on low shot k-NN)

Thank you

What to Hide from Your Students: Attention-Guided Masked Image Modeling

Wednesday 26/10

ECCV Posters Session 2.B

Paper ID #1439

Posterboard #77

Paper: https://arxiv.org/abs/2203.12719

Code: https://github.com/gkakogeorgiou/attmask