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

Ioannis Kakogeorgiou[1], Spyros Gidaris[2], Bill Psomas[1], Yannis Avrithis[3,4], Andrei Bursuc[2], Konstantinos Karantzalos[1], Nikos Komodakis[5,6]

[1]National Technical University of Athens; [2]valeo.ai; [3]Institute of Advanced Research in Artificial Intelligence (IARAI); [4]Athena RC; [5]University of Crete; [6] IACM-Forth

gkakogeorgiou@central.ntua.gr

ECCV TEL AVIV 2022

Introduction

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

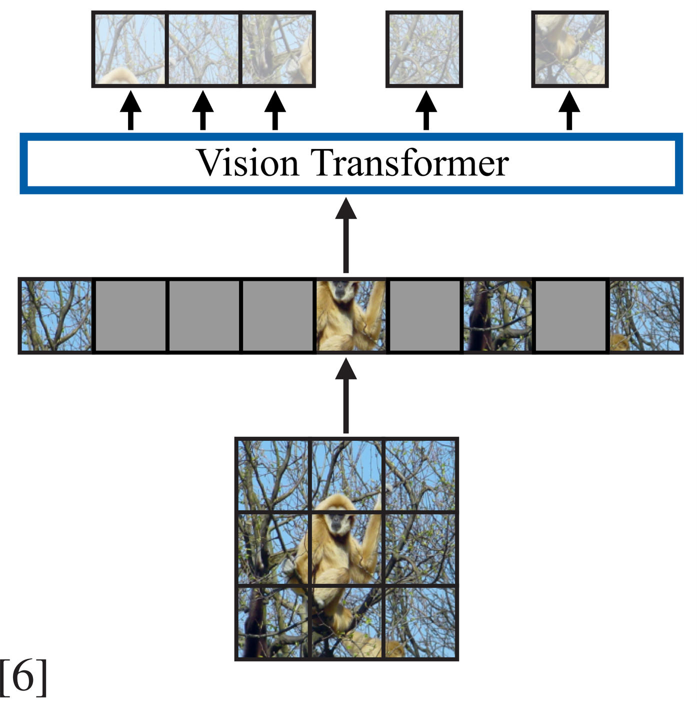
- Mask a portion of the input patch tokens
- Train a Transformer to reconstruct them

Focus: Which patch tokens to mask?

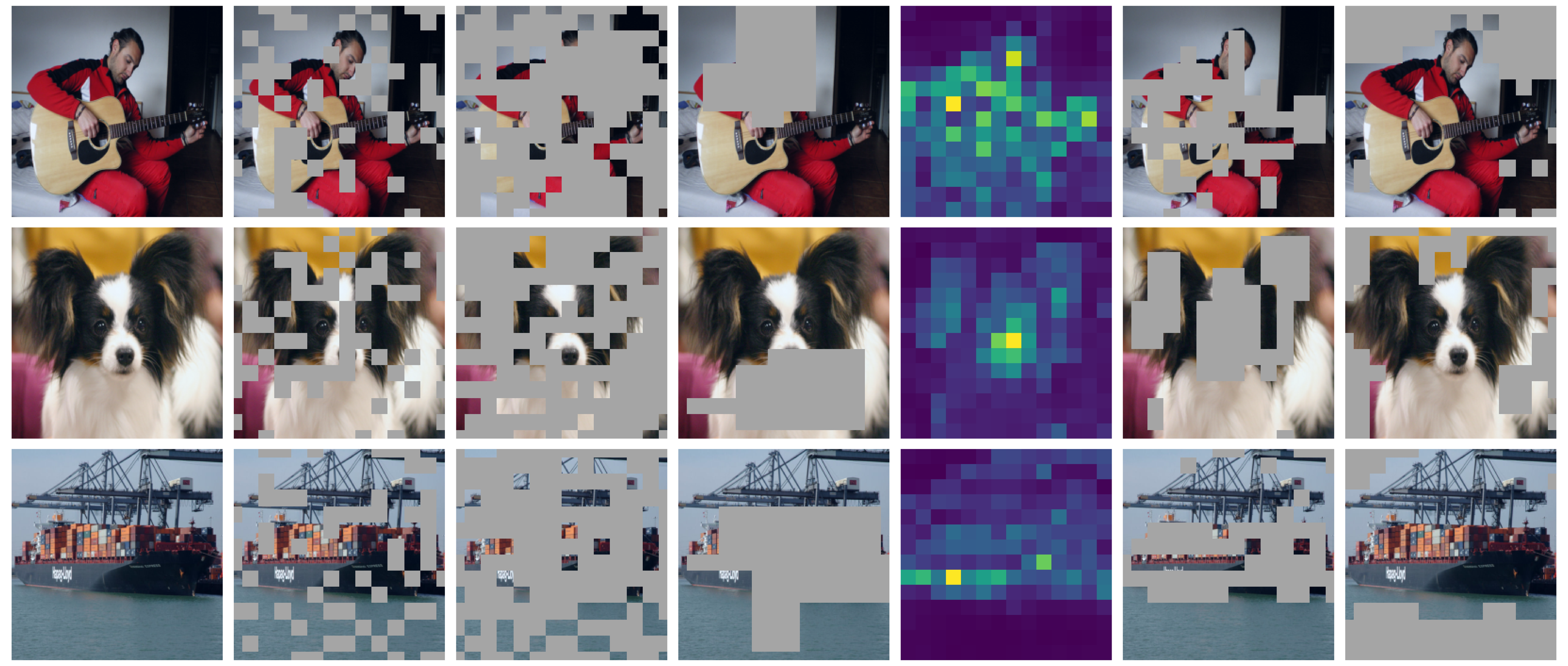
- Not well explored
- Prior works use (block-wise) random token masking

Approach: Attention-guided token masking (AttMask)

- Leverage ViT's self-attention to mask highly-attended tokens
- Excellent fit to distillation-based approaches, e.g., iBOT [1], DINO [6]



AttMask: Attention-Guided Masked Image Modeling



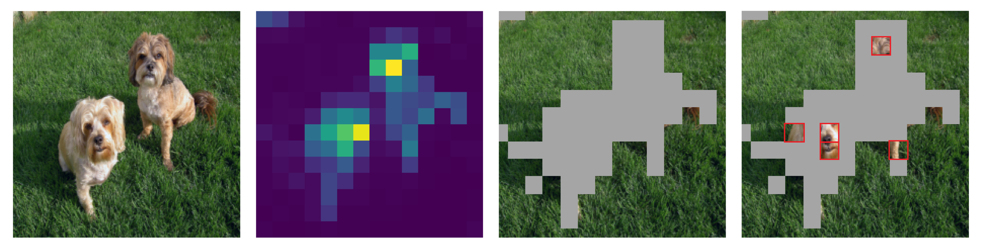
Input Image | Random (30%) [2] | Random (75%) [4] | Block Wise [3] | Attention Map | AttMask-High | AttMask-Low

Issues with (block-wise) random masking

- Less likely to hide "interesting" parts → easy reconstruction
- Compensating with extreme masking (e.g., 75% of tokens) → overly aggressive

Exploring attention-guided masking (AttMask):

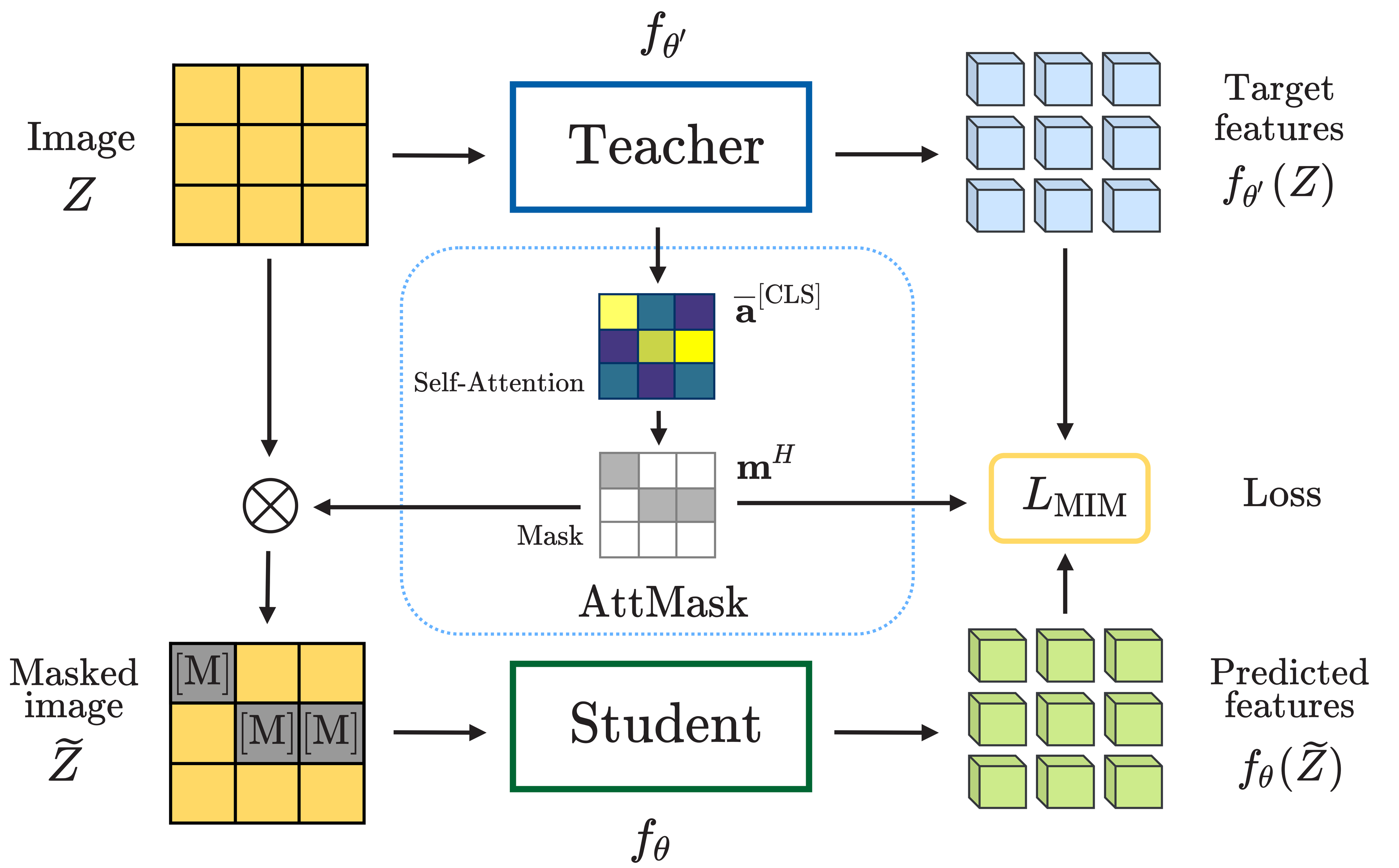
AttMask	Masked Tokens	Task	Performance
✗ Low	low-attended	very easy	⇓
✓ High	high-attended	very challenging	⇑
✓ Hint	high-attended, except hints	challenging	⇑



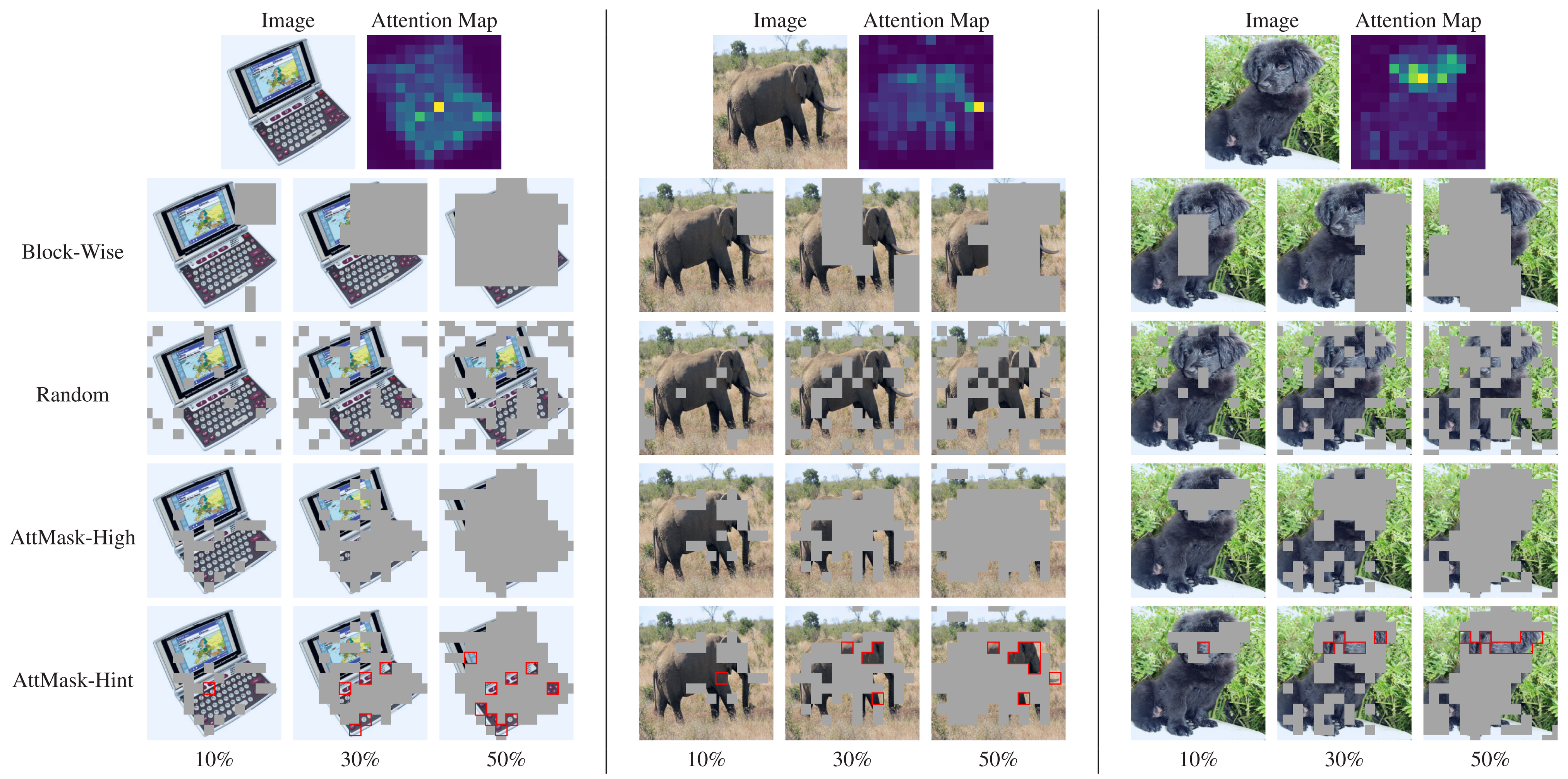
Input Image | Attention Map | AttMask-High | AttMask-Hint

Incorporating AttMask into distillation-based methods

Attention map from: [CLS] token in the last self-attention layer of the teacher



Qualitative examination of masking strategies

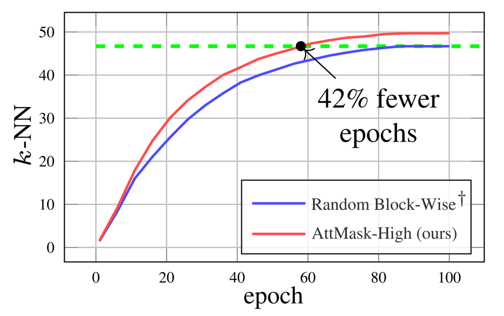


Experimental Results

Incorporating AttMask into the MIM-based self-supervised method iBOT [1] using ViT

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

iBOT Masking	Ratio (%)	ImageNet-1k k-NN	ImageNet-1k Linear	CIFAR10 Fine-tuning	CIFAR100 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**



Mask Ratio r (%)	10-30	10-50	10-70	30
Random Block-Wise	46.5	46.7†	47.1	46.9
Random	47.6	47.8	47.8	48.2
AttMask-High	49.5	**49.7**	48.5	49.1

Evaluating on ImageNet-1k by pre-training on full ImageNet-1k for 100 (left) and 300 (right) epochs

Method	Full k-NN	Full Linear	Few examples $\nu=1$	5	10	20
DINO [6]	70.9	74.6				
MST [5]	72.1	75.0				
iBOT [1]	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**

Method	Full k-NN	Full Linear	Few examples $\nu=1$	5	10	20
SimCLR [7]	-	69.0				
BYOL [8]	66.6	71.4				
MoBY [9]	-	72.8				
DINO [6]	72.8	76.1				
MST [5]	**75.0**	76.9				
iBOT [1]	74.6	77.4	38.9	54.1	58.5	61.9
iBOT+AttMask-High	**75.0**	**77.5**	**40.4**	**55.5**	**59.9**	**63.1**

Transfer learning with fine-tuning on object detection (COCO) and semantic segmentation (ADE20K) and without fine-tuning on Image Retrieval ($\mathcal{R}$Oxford and $\mathcal{R}$Paris) and video object segmentation (DAVIS).

Method	COCO AP^b	COCO AP^m	ADE20K mIoU	$\mathcal{R}$Oxford Medium	$\mathcal{R}$Oxford Hard	$\mathcal{R}$Paris Medium	$\mathcal{R}$Paris Hard	DAVIS 2017 $(\mathcal{J}\&\mathcal{F})_m$	DAVIS 2017 $\mathcal{J}_m$	DAVIS 2017 $\mathcal{F}_m$
iBOT	48.2	41.8	44.9	31.0	11.7	56.2	28.9	60.5	59.5	61.4
iBOT+AttMask	**48.8**	**42.0**	**45.3**	**33.5**	**12.1**	**59.0**	**31.5**	**62.1**	**60.6**	**63.5**

Conclusion

- Zero additional cost
- Benefits over random masking
- Outperforms the other self-supervised distillation-based MIM methods
- Major improvements in challenging tasks; i.e., using features without additional finetuning, or working with limited data.

References

[1] Zhou et al. iBOT: Image BERT Pre-training with Online Tokenizer *ICLR*, 2022.
[2] Xie et al. SimMIM: A Simple Framework for Masked Image Modeling *CVPR*, 2022.
[3] Bao et al. BEiT: BERT Pre-Training of Image Transformers *ICLR*, 2022.
[4] He et al. Masked Autoencoders Are Scalable Vision Learners *CVPR*, 2022.
[5] Li et al. MST: Masked Self-Supervised Transformer for Visual Representation *NIPS*, 2021.
[6] Caron et al. Emerging properties in self-supervised vision transformers *ICCV*, 2021.
[7] Chen et al. A simple framework for contrastive learning of visual representations *ICML*, 2020.
[8] Grill et al. Bootstrap your own latent-a new approach to self-supervised learning *NIPS*, 2020.
[9] Xie et al. Self-supervised learning with swin transformers *arXiv*, 2021.