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

Input Image → Encoder → Features
Representation Learning

Input Image → Encoder → Features → Global Representation
Representation Learning

Input Image

Encoder

Features

Global Representation

“fish”

Image Classification

Image Retrieval
Representation Learning

Input Image → Encoder → Features → Local Representations
Representation Learning

Input Image

Encoder

Features

Local Representations

Object Detection

Semantic Segmentation
How to learn the Encoder?

Input Image

Encoder

Features

Global Representation

Local Representations

“fish”
- Image Classification
- Image Retrieval
- Object Detection
- Semantic Segmentation
How to learn the Encoder: from scratch

Input Image

• Supervised, from scratch, for each task separately

Encoder

Global Representation

"fish"
Image Classification

Image Retrieval

Object Detection

Semantic Segmentation

Local Representations
How to learn the Encoder: from scratch

- **Supervised, from scratch, for each task separately**
  - Labor-intensive...

- “fish”
  - Image Classification
  - Image Retrieval
  - Object Detection
  - Semantic Segmentation
How to learn the Encoder: Transfer Learning

- Supervised, two stage: firstly learn on classification (cheap)
How to learn the Encoder: Transfer Learning

- Supervised, two stage: firstly learn on classification (cheap) and then downstream to other tasks
  ✗ Better, but still labor-intensive...

Local Representations

Global Representation

“holoacanthus” Image Classification
Image Retrieval
Object Detection
Semantic Segmentation
How to learn the Encoder: Self-supervised Learning

- Self-supervised, two stage: firstly, learn on a pretext task (free)

Input Image → Encoder → Features → Global Representation → “90°”

Pretext Task: Image Rotation
How to learn the Encoder: Self-supervised Learning

- Self-supervised, two stage: firstly, learn on a pretext task (free) and then downstream to other tasks

✓ Best, pre-training labels are automatically generated!
Self-supervised pretext tasks

1. Solving the pretext tasks allow the model to learn **good features**
2. We can **automatically** generate **labels** for the pretext tasks
Self-supervised pretext tasks

rotation prediction  “jigsaw puzzle”  colorization

✗ Learned representations may be tied to a specific pretext task!

Can we come up with a more general pretext task?
A more general pretext task?

same subject
A more general pretext task?

same subject

different subject

Stanford University CS231n: Deep Learning for Computer Vision
Self-supervised Contrastive Learning

attract

repel
Leveraging Attention in Masked Image Modeling
Masked Image Modeling (MIM)

- Divide an input image into patch tokens
Masked Image Modeling (MIM)

- Divide an input image into patch tokens
- Mask a portion of the input patch tokens
Masked Image Modeling (MIM)

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

Zhou et al., iBOT: Image BERT Pre-training with Online Tokenizer ICLR, 2022
Bao et al., BEiT: BERT Pre-Training of Image Transformers ICLR, 2022
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
Approach: Attention-guided token masking (AttMask)

- Leverage ViT’s self-attention to mask tokens
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
Leverage ViT’s self-attention to mask tokens

✓ **AttMask-High**: masks highly-attended tokens (essentially foreground)

→ very challenging reconstruction task → boosts performance
Approach: Attention-guided token masking (AttMask)

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

Perhaps overly aggressive for high mask ratios!
Approach: Attention-guided token masking (AttMask)

- Leverage ViT’s self-attention to mask tokens
  ✓ **AttMask-High**: masks highly-attended tokens (essentially foreground)
    \[\rightarrow\text{very challenging reconstruction task} \rightarrow \text{boosts performance}\]
  ✓ **AttMask-Hint**: masks highly-attended tokens, but leaves some hints
    \[\rightarrow\text{provides hints for the identity of the masked object} \rightarrow \text{boosts performance}\]

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Incorporating AttMask into distillation-based methods

We exhibit AttMask in the context of distillation-based MIM, such as iBOT.
Incorporating AttMask into distillation-based methods

- We exhibit AttMask in the context of distillation-based MIM, such as iBOT
- The teacher transformer encoder sees the entire image and generates the attention map

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Incorporating AttMask into distillation-based methods

- We exhibit AttMask in the context of distillation-based MIM, such as iBOT
- The teacher transformer encoder sees the entire image and generates the attention map

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Incorporating AttMask into distillation-based methods

- We exhibit AttMask in the context of distillation-based MIM, such as iBOT
- The teacher transformer encoder sees the entire image and generates the attention map

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Incorporating AttMask into distillation-based methods

- We exhibit AttMask in the context of distillation-based MIM, such as iBOT
- 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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Incorporating AttMask into distillation-based methods

- We exhibit AttMask in the context of **distillation-based MIM**, such as iBOT
- 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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Incorporating AttMask into distillation-based methods

- We exhibit AttMask in the context of distillation-based MIM, such as iBOT
- 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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Qualitative examination of masking strategies

- Input image
- Attention map

**block-wise**
- Random
- AttMask High
- AttMask Hint

10% 30% 50%

---

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Evaluating token masking strategies (20% of ImageNet-1k)

| iBOT MASKING          | RATIO (%) | IMAGE NET-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 |

Top-1 accuracy for k-NN and linear probing

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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Evaluating token masking strategies (20% of ImageNet-1k)

| 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** |

*Top-1 accuracy for k-NN and linear probing*

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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Evaluating token masking strategies (20% of ImageNet-1k)

| 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** |

Top-1 accuracy for k-NN and linear probing

- AttMask-High **improves iBOT** by +3% on k-NN and +1.5% on linear probing
Evaluating token masking strategies (20% of ImageNet-1k)

| IBOT MASKING       | RATIO (%) | IMAGE NET-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     |

Top-1 accuracy for k-NN and linear probing

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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Evaluating token masking strategies (different % 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 k-NN accuracy for pre-training on different percentages of ImageNet-1k

Improved performance when:
- Pre-training with fewer data

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Evaluating token masking strategies (different % of ImageNet-1k)

†: default iBOT masking strategy from BEiT

| % 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** |

Top-1 k-NN accuracy for pre-training on different percentages of ImageNet-1k

| METHOD                 | FULL | FEW EXAMPLES |
|------------------------|------|--------------|
|                         | k-NN | ν = 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** |

Top-1 accuracy for pre-training on 100% of ImageNet-1k
(a) k-NN and linear probing
(b) k-NN using only few examples per class

Improved performance when:
✓ Pre-training with fewer data
✓ Pre-training on the full ImageNet-1k (+1.3% on k-NN and +1.5% on linear probing)

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Property: Low-shot performance

\[\text{\textsuperscript{\dag}}: \text{default iBOT masking strategy from BEiT}\]

\[
\begin{array}{l|cccc}
\% \ \text{IMAGENET-1K} & 5 & 10 & 20 & 100 \\
\hline
\text{Random Block-Wise}^{\text{\dag}} & 15.7 & 31.9 & 46.7 & 71.5 \\
\text{AttMask-High (ours)} & \textbf{17.5} & \textbf{33.8} & \textbf{49.7} & \textbf{72.5} \\
\end{array}
\]

Top-1 k-NN accuracy for pre-training on different percentages of ImageNet-1k

Improved performance when:

- Pre-training with fewer data
- Pre-training 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)

| Method                     | Full $k$-NN | Linear $\nu = 1$ | Few Examples |
|----------------------------|-------------|------------------|--------------|
| DINO                       | 70.9        | 74.6             |              |
| MST                        | 72.1        | 75.0             |              |
| iBOT                       | 71.5        | 74.4             | 32.9         |
| iBOT+AttMask-High          | 72.5        | 75.7             | 37.1         |
| iBOT+AttMask-Hint          | \textbf{72.8} | \textbf{76.1} | \textbf{37.6} |

(a) k-NN and linear probing
(b) k-NN using only few examples per class

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
### Property: Background robustness

| IBOT MASKING         | RATIO (%) | OF | MS | MR | MN | NF | OBB | OBT | IN-9 |
|----------------------|-----------|----|----|----|----|----|-----|-----|------|
| Random Block-wise †  | 10-50     | 72.4 | 74.3 | 59.4 | 56.8 | 36.3 | 14.4 | 15.0 | 89.1 |
| Random ‡             | 75        | 73.1 | 73.8 | 58.8 | 55.9 | 35.6 | 13.7 | 14.5 | 87.9 |
| Random               | 10-50     | 72.8 | 75.3 | 60.4 | 57.5 | 34.9 | 10.3 | 14.4 | 89.3 |
| AttMask-Low (ours)   | 10-50     | 66.0 | 71.1 | 55.2 | 52.2 | 32.4 | 12.5 | 14.0 | 86.6 |
| AttMask-Hint (ours)  | 10-50     | 74.4 | 75.9 | 61.7 | 58.3 | 39.6 | **16.7** | **15.7** | 89.6 |
| AttMask-High (ours)  | 10-50     | **75.2** | **76.2** | **62.3** | **59.4** | **40.6** | 15.2 | 15.3 | **89.8** |

Classification **robustness** against **background changes**

Classification accuracy of linear probe on IN-9 and its variations

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Downstream tasks

Object detection (COCO) and semantic segmentation (ADE20K) with fine-tuning
Image Retrieval (ROXFORD and RPARIS) and video object segmentation (DAVIS) without fine-tuning

- Improved performance on downstream tasks with or without fine-tuning
Property: High-quality features

| METHOD       | COCO | ADE20K | ROXFORD | RPARIS | DAVIS 2017 |
|--------------|------|--------|---------|--------|-----------|
|              | AP^b | AP^m  | mIoU    | MEDIUM | HARD      | MEDIUM | HARD | (J & F)_m | J_m | 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      |

Object detection (COCO) and semantic segmentation (ADE20K) with fine-tuning
Image Retrieval (ROXFORD and RPARIS) and video object segmentation (DAVIS) without fine-tuning

✓ Improved performance on downstream tasks with or without fine-tuning

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Conclusion

AttMask:

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

Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022
Collaborators

Ioannis Kakogeorgiou
Spyros Gidaris
Andrei Bursuc
Konstantinos Karantzalos
Yannis Avrithis
Nikos Komodakis