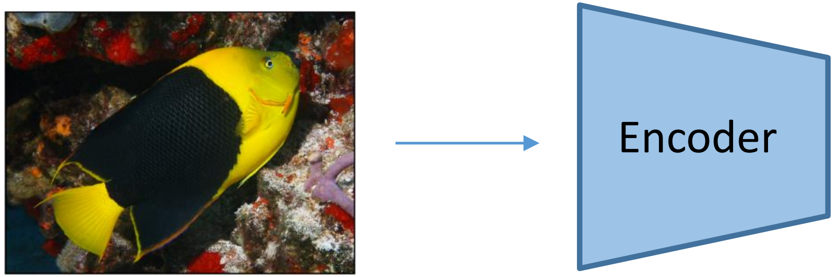
# Leveraging Attention in Masked Image Modeling and Pooling

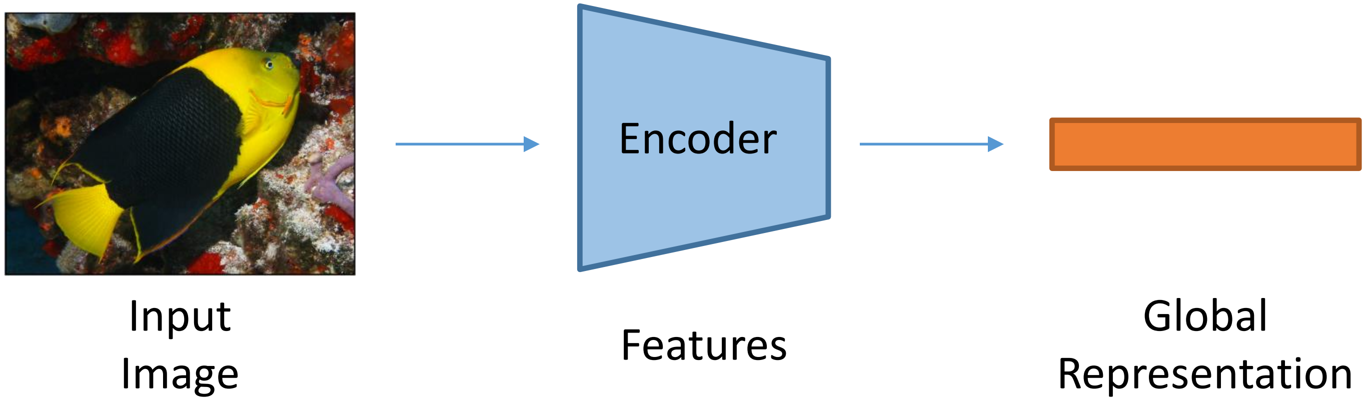
# Representation Learning



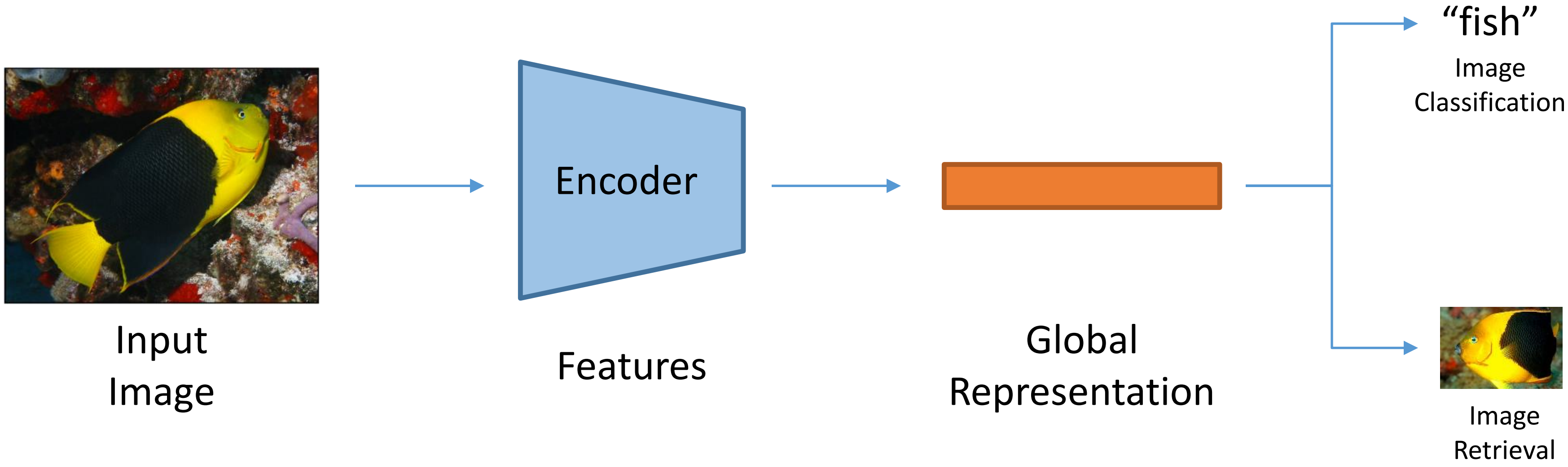

Input
Image

Features

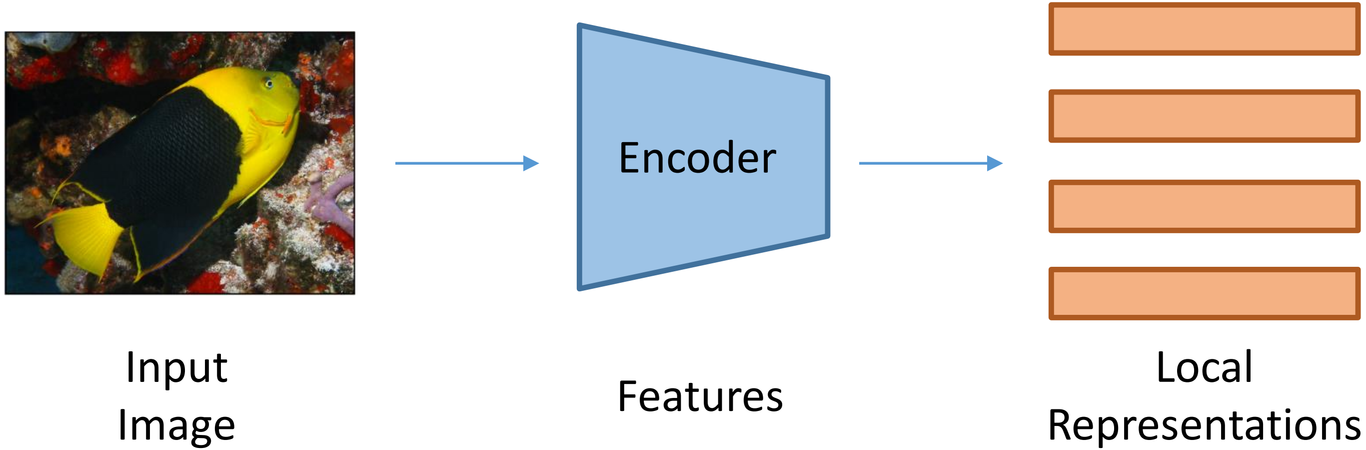
# Representation Learning



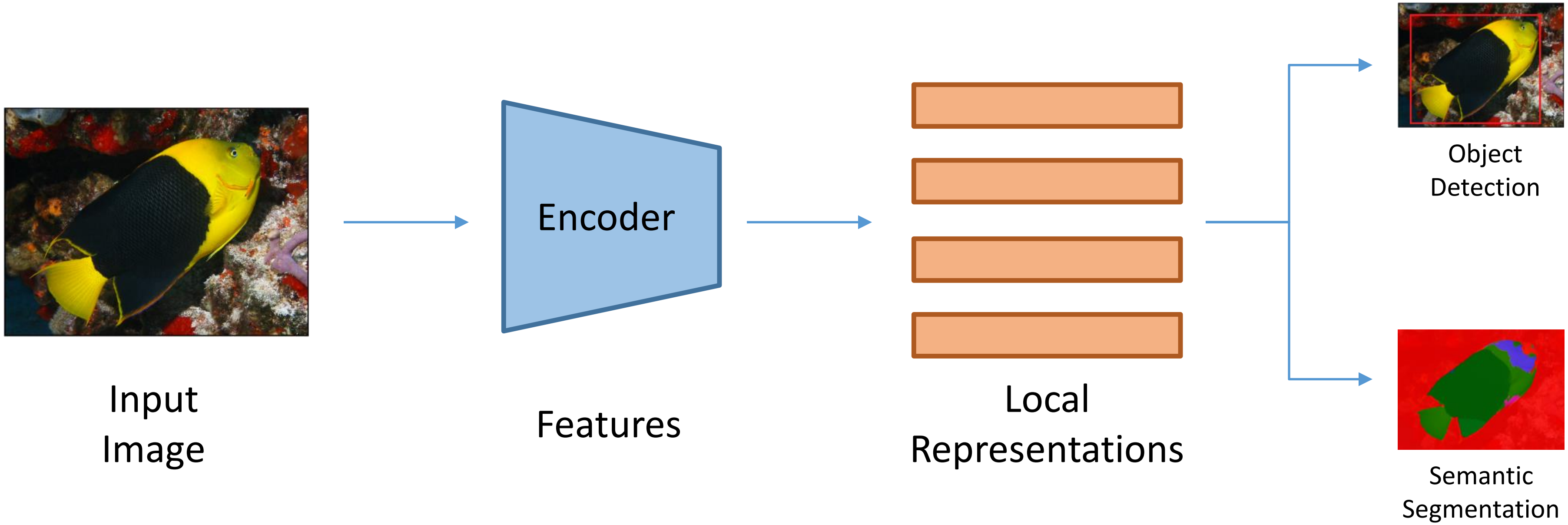
# Representation Learning



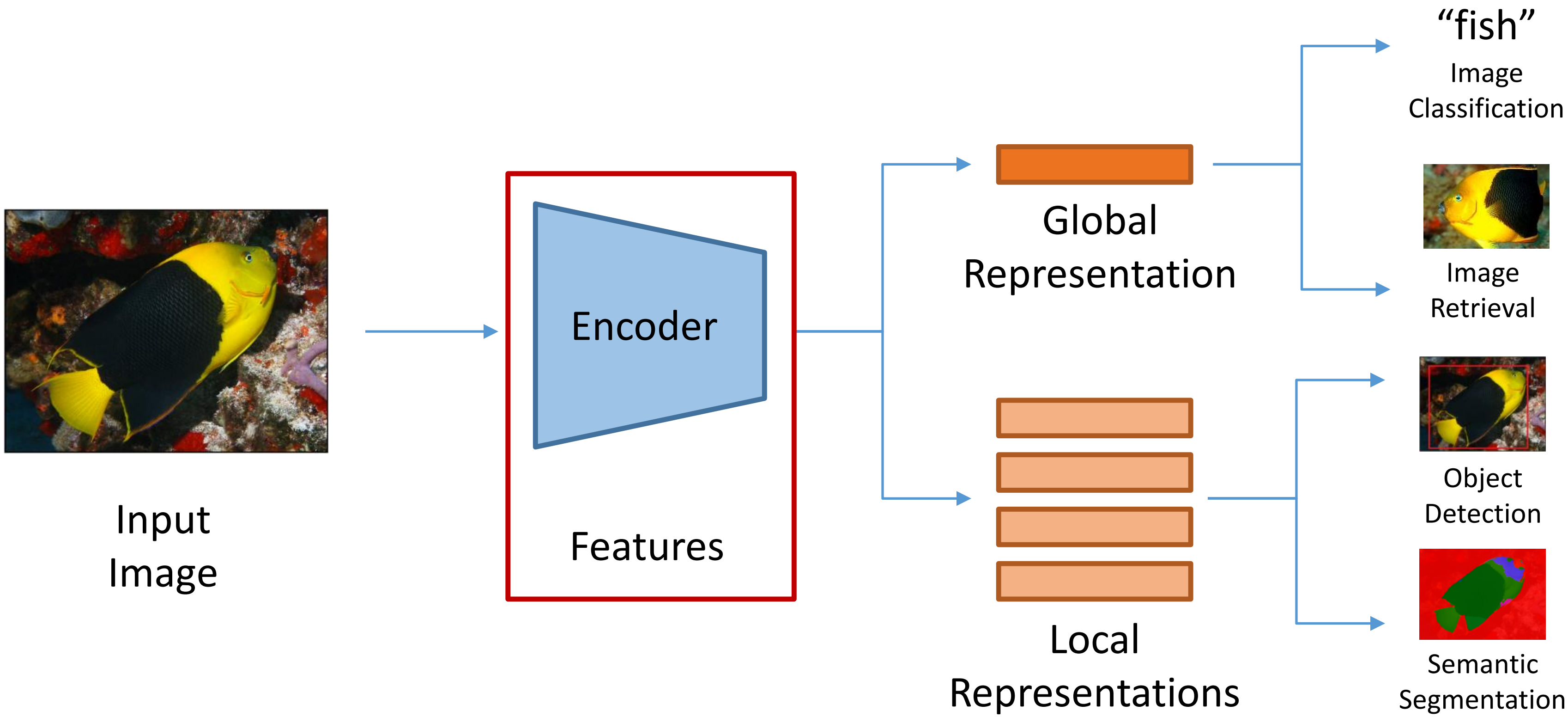
# Representation Learning



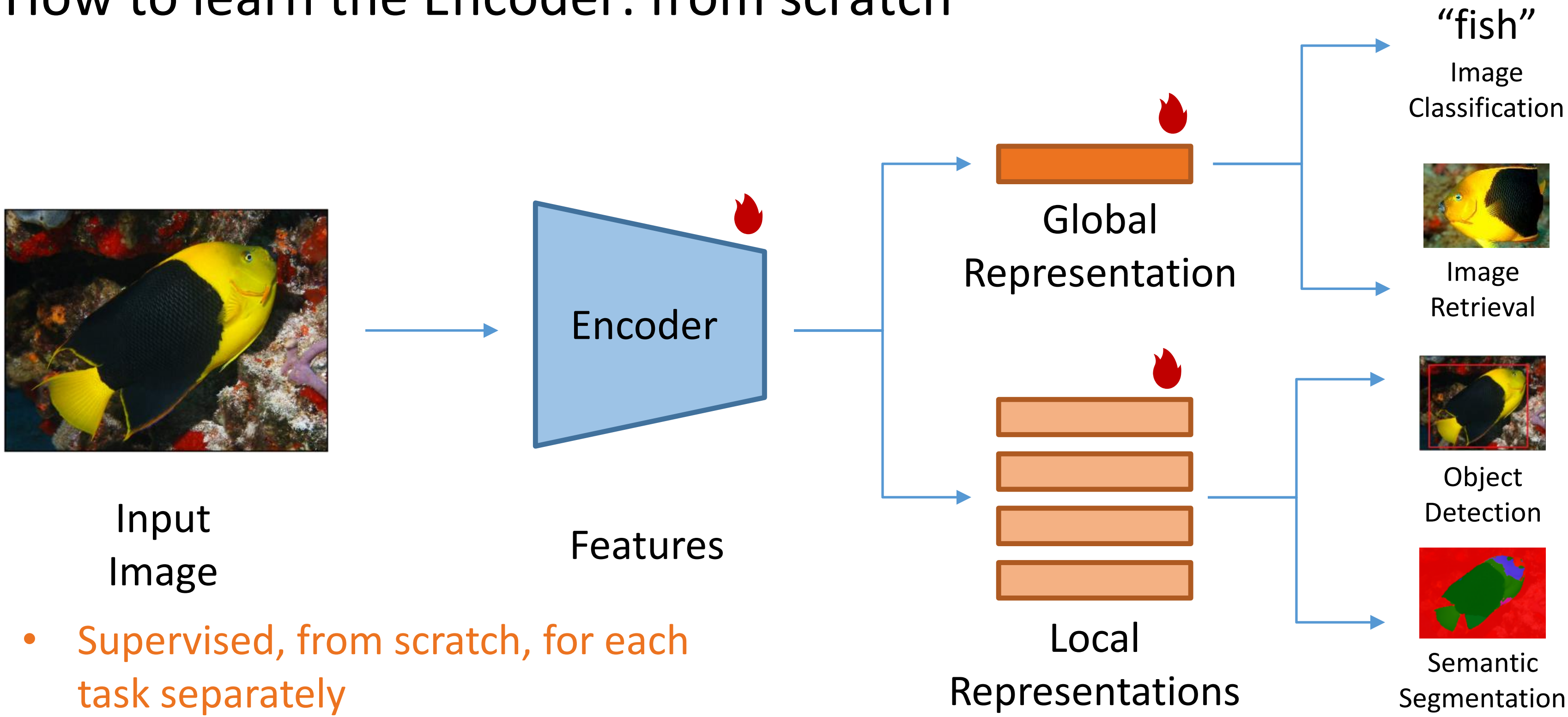
# Representation Learning



# How to learn the Encoder?

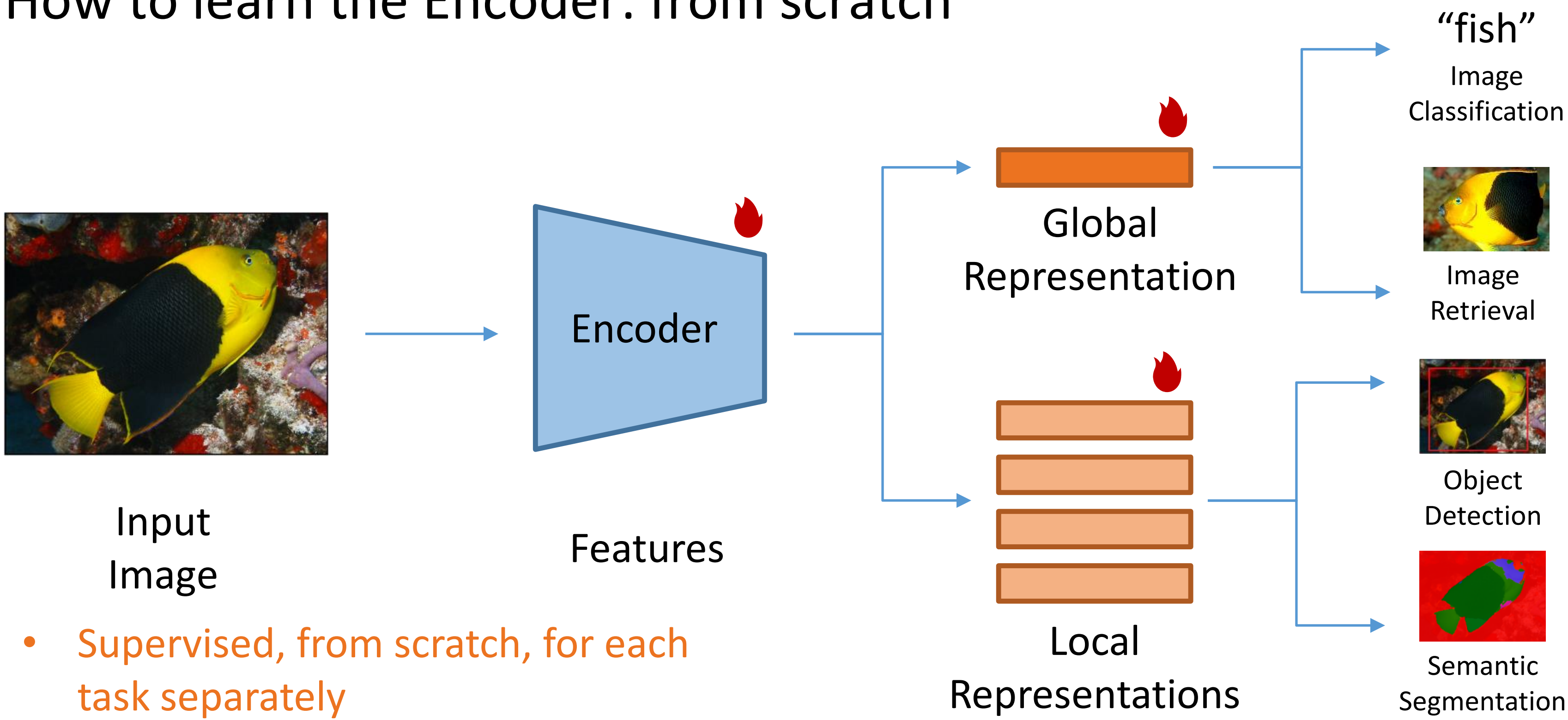


# How to learn the Encoder: from scratch




- Supervised, from scratch, for each task separately

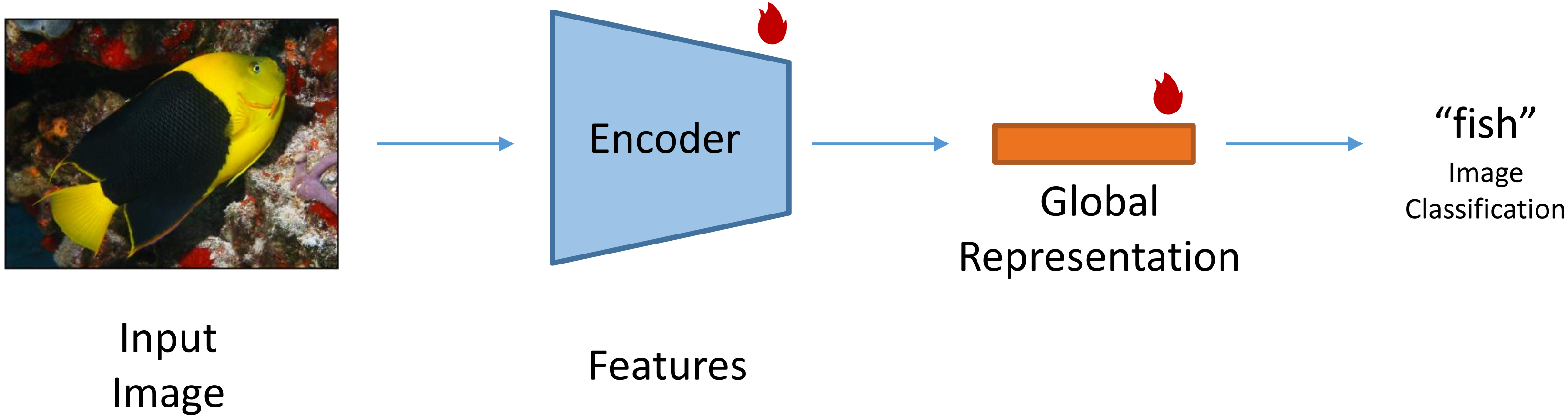
# How to learn the Encoder: from scratch




- Supervised, from scratch, for each task separately

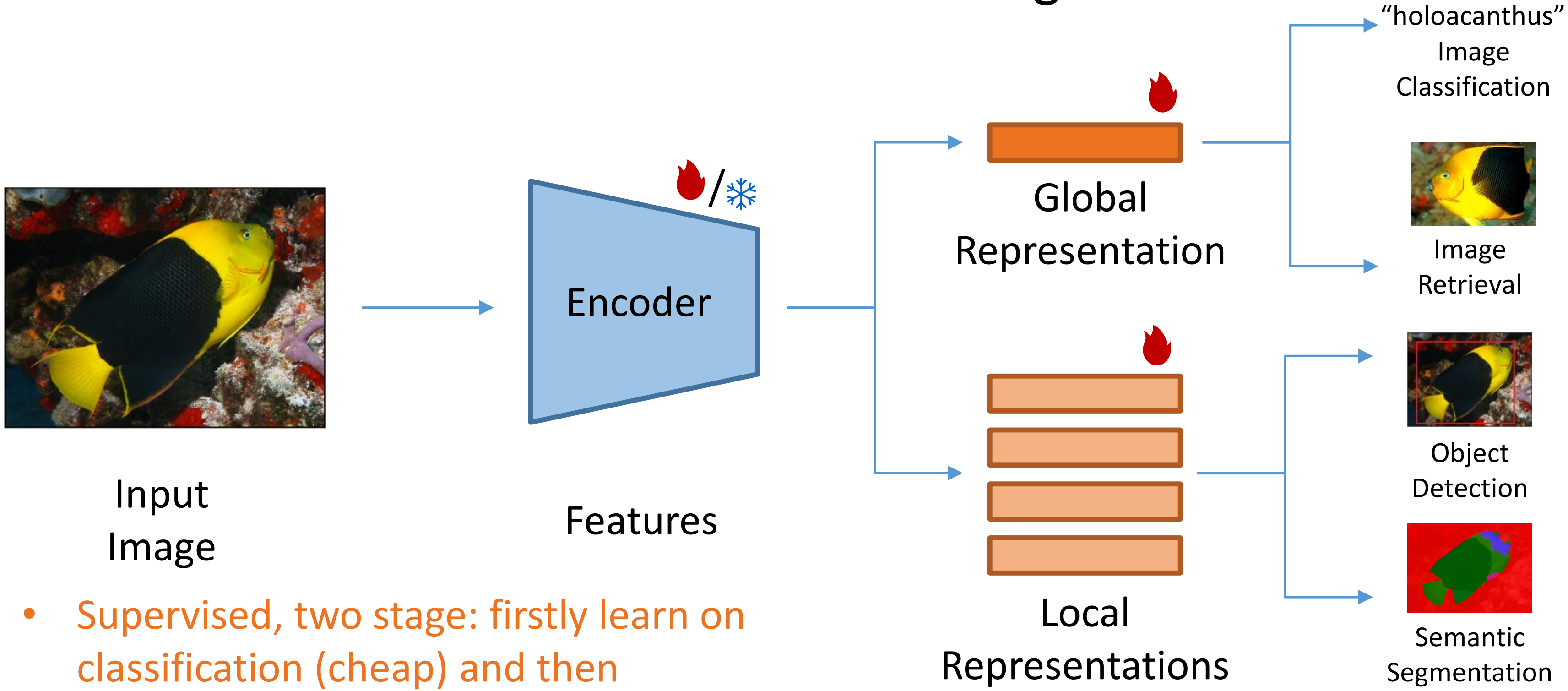
  ✘ Labor-intensive…

# How to learn the Encoder: Transfer Learning



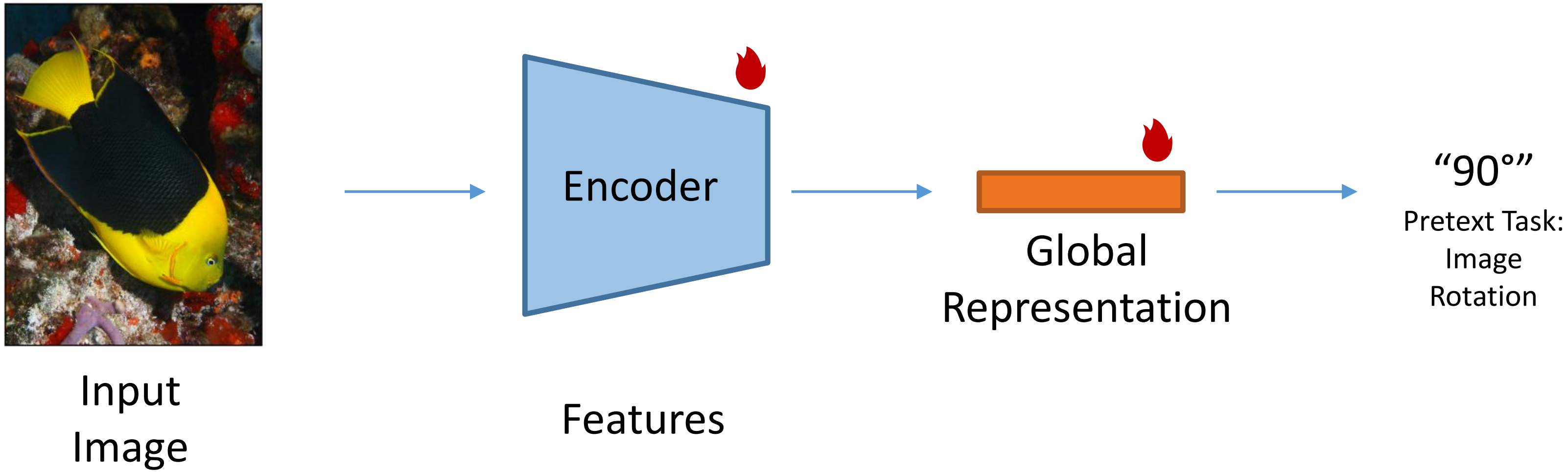

- Supervised, two stage: firstly learn on classification (cheap)

# How to learn the Encoder: Transfer Learning



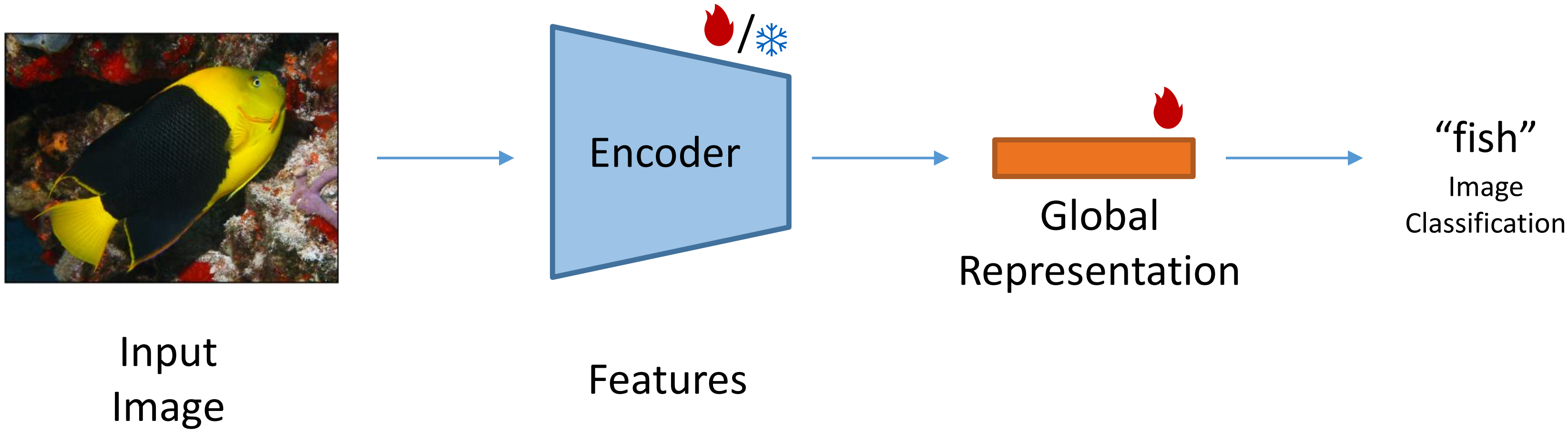

- Supervised, two stage: firstly learn on classification (cheap) and then downstream to other tasks
  - ✘ Better, but still labor-intensive…

# How to learn the Encoder: Self-supervised Learning




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

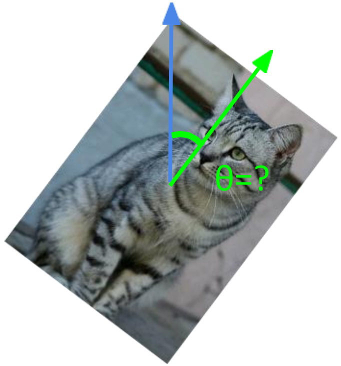
# How to learn the Encoder: Self-supervised Learning




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

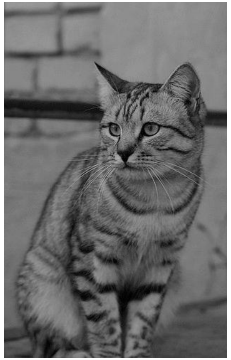
✓Best, pre-training labels are automatically generated!

# Self-supervised pretext tasks




rotation prediction

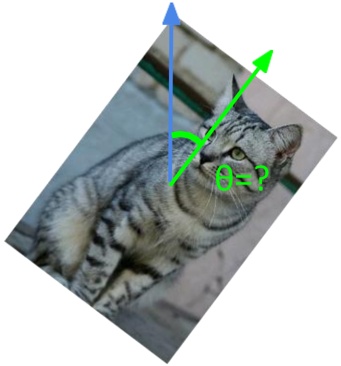
“jigsaw puzzle”



colorization

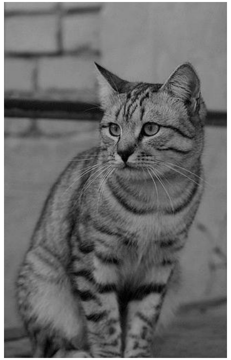
1. Solving the pretext tasks allow the model to learn good features
2. We can automatically generate labels for the pretext tasks

# Self-supervised pretext tasks




rotation prediction

"jigsaw puzzle"

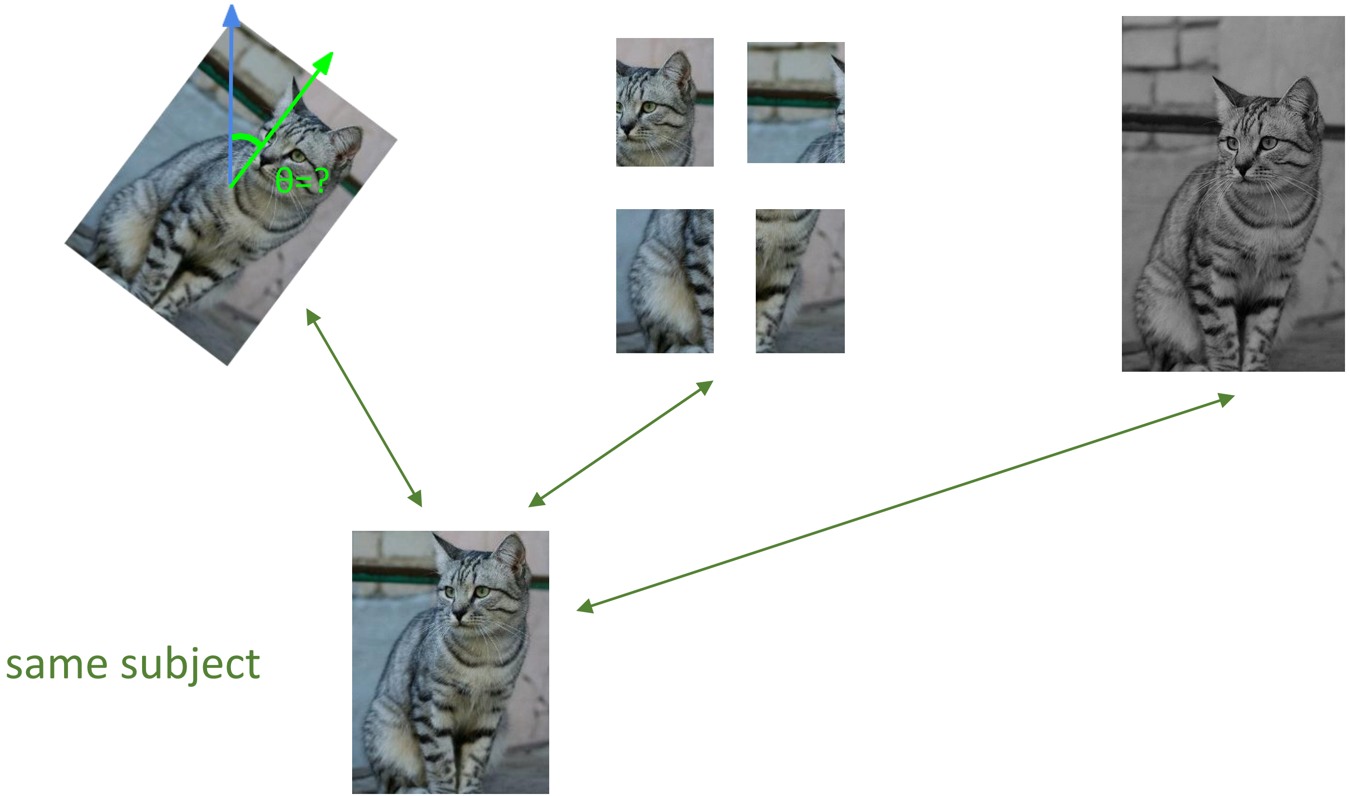


colorization

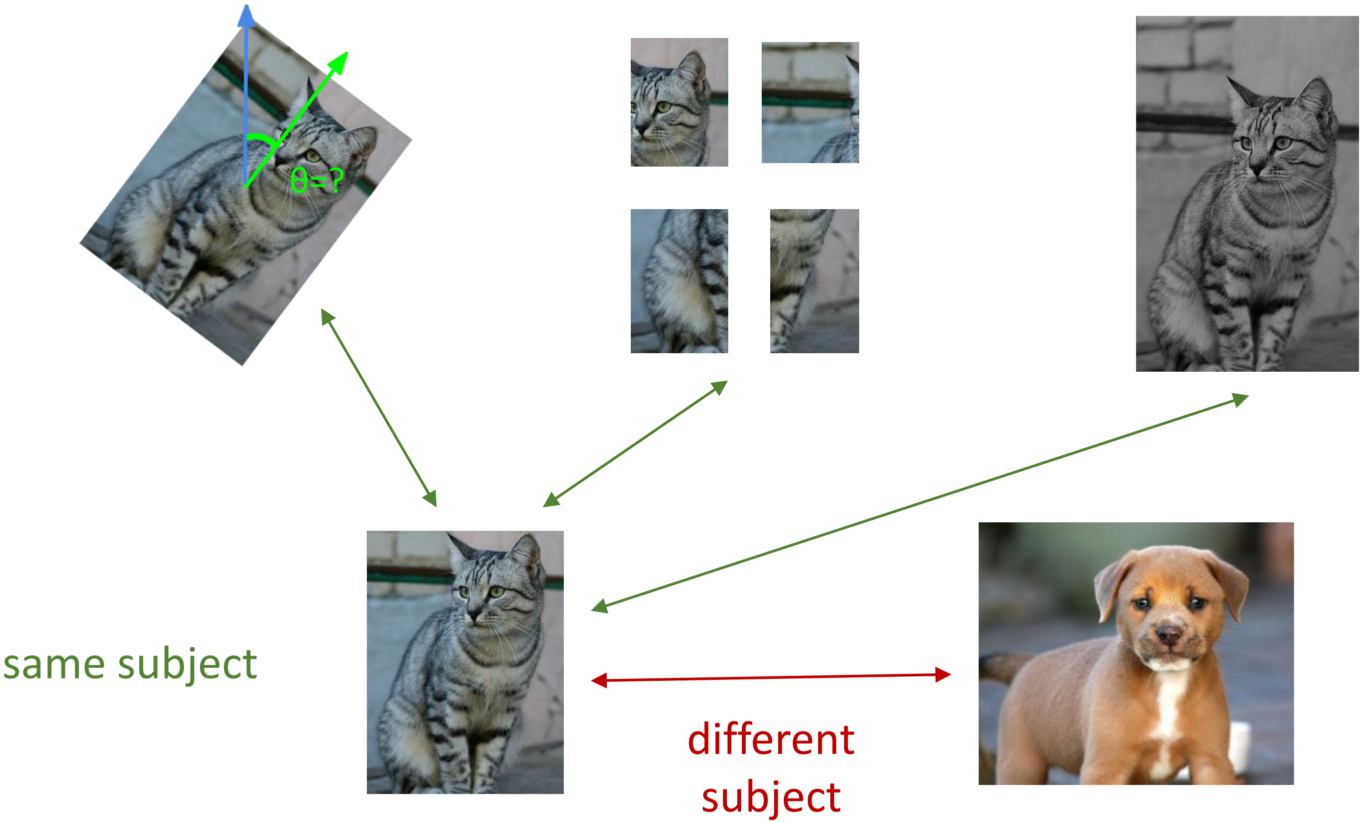
✘ Learned representations may be tied to a specific pretext task!

Can we come up with a more general pretext task?

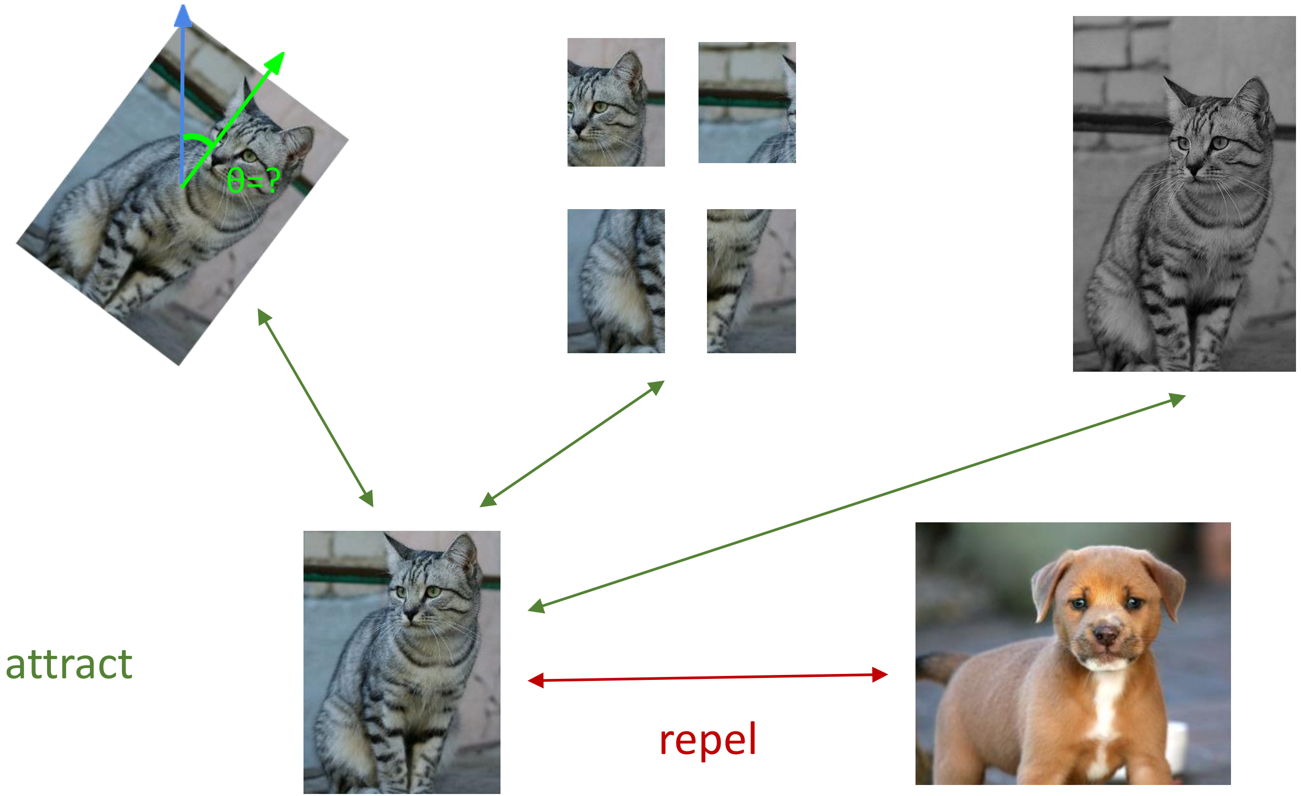
# A more general pretext task?



# A more general pretext task?



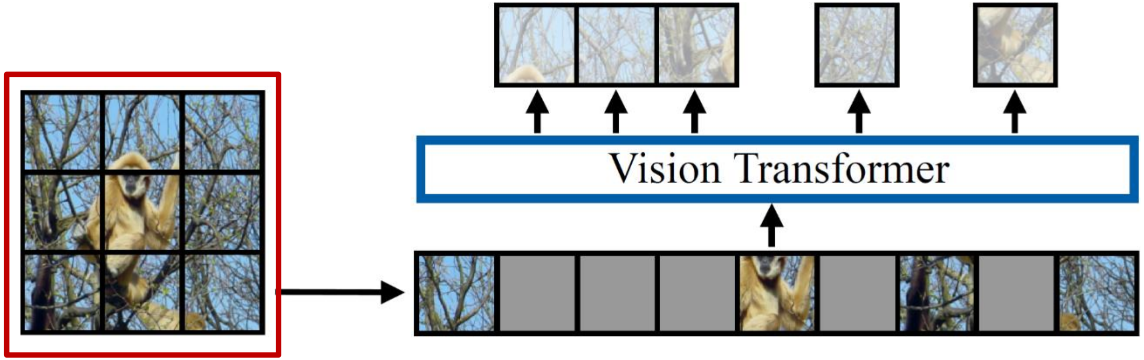
# Self-supervised Contrastive Learning




Stanford University CS231n: Deep Learning for Computer Vision

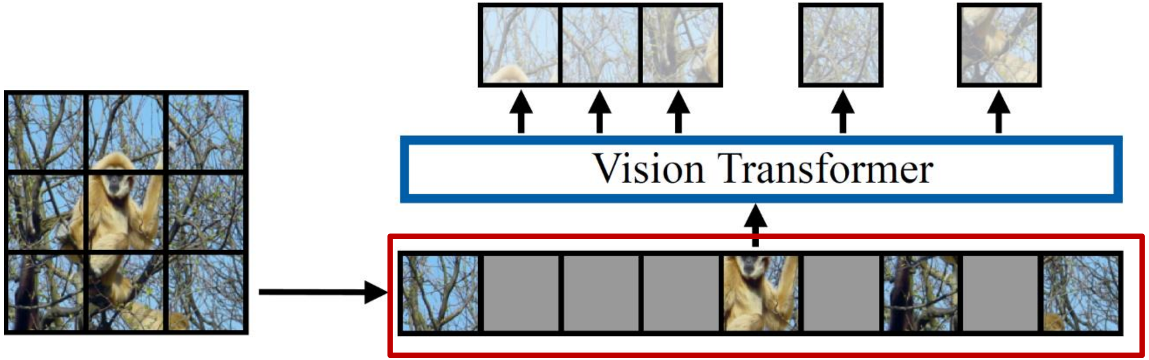
# Leveraging Attention in Masked Image Modeling

# Masked Image Modeling (MIM)



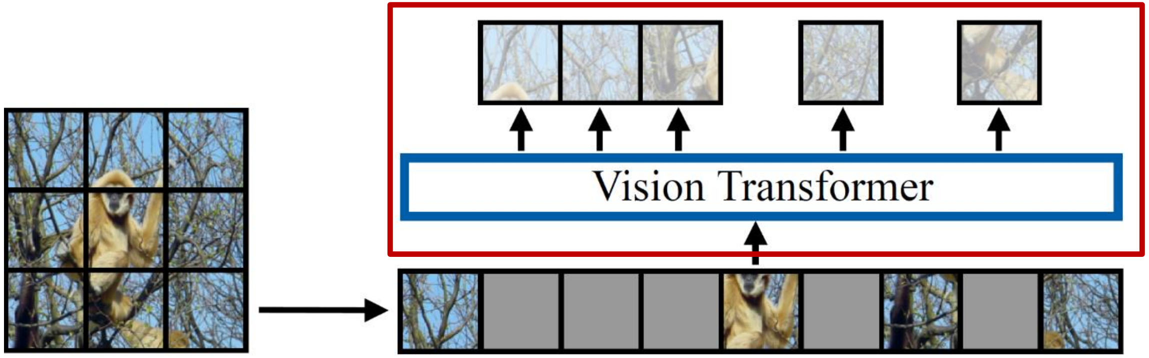

- Divide an input image into patch tokens

# 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

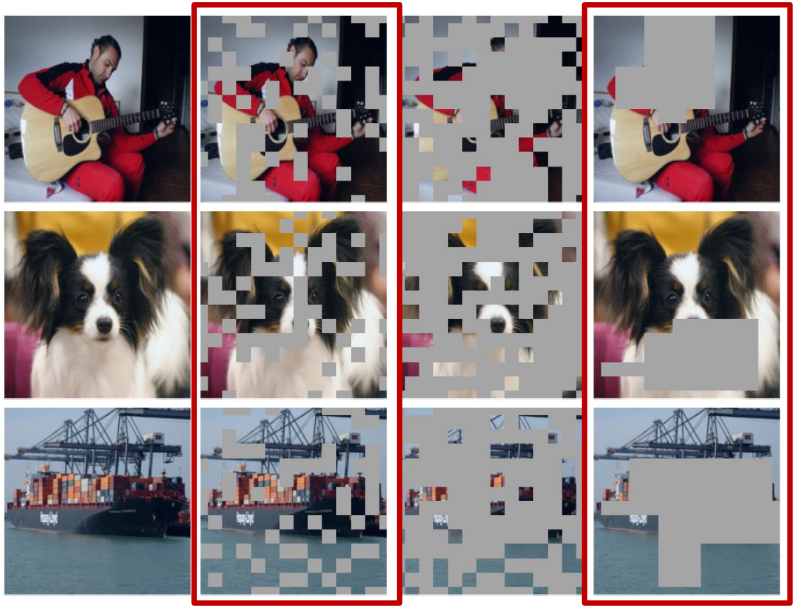
# Focus: Which patch tokens to mask?

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

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

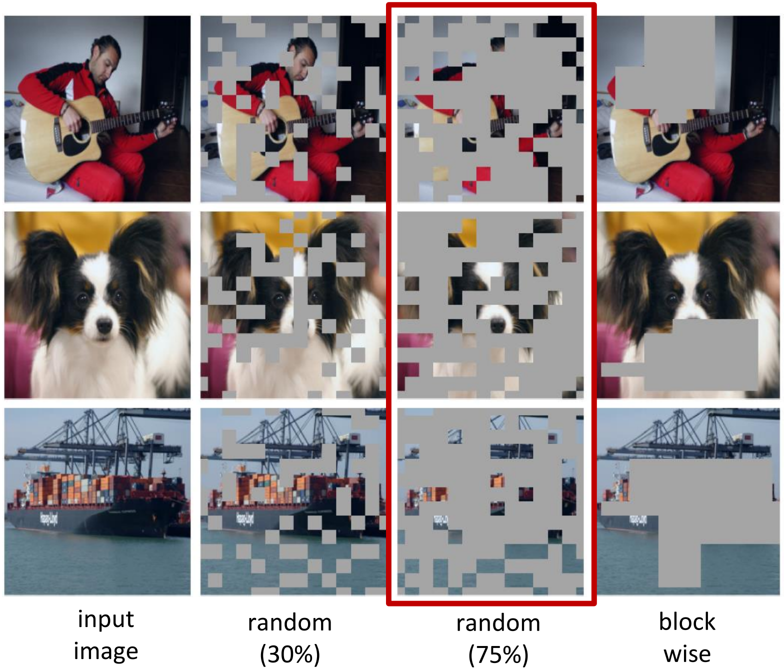


input image | random (30%) | random (75%) | block wise

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?

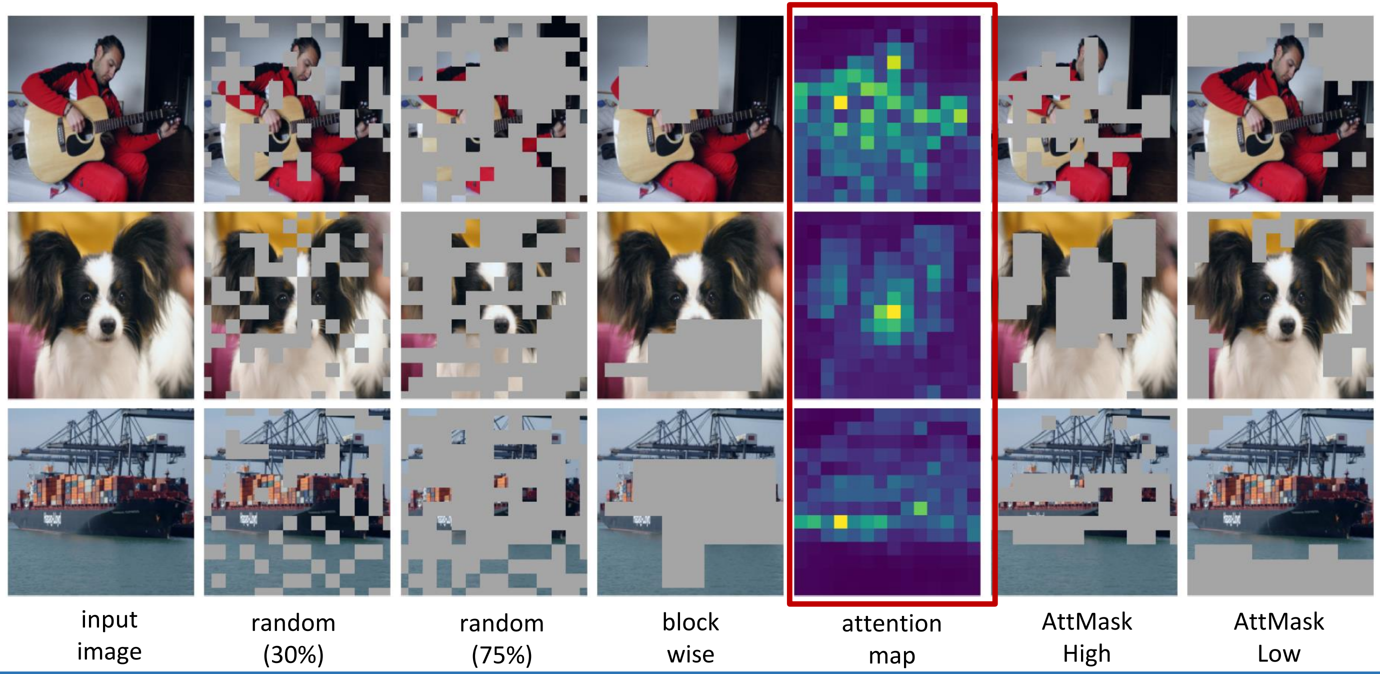
- Not well explored; prior works use (block-wise) random token masking
  - Less likely to hide "interesting" parts → easy reconstruction
  - Compensating with extreme masking (e.g. 75% of tokens) → overly aggressive




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

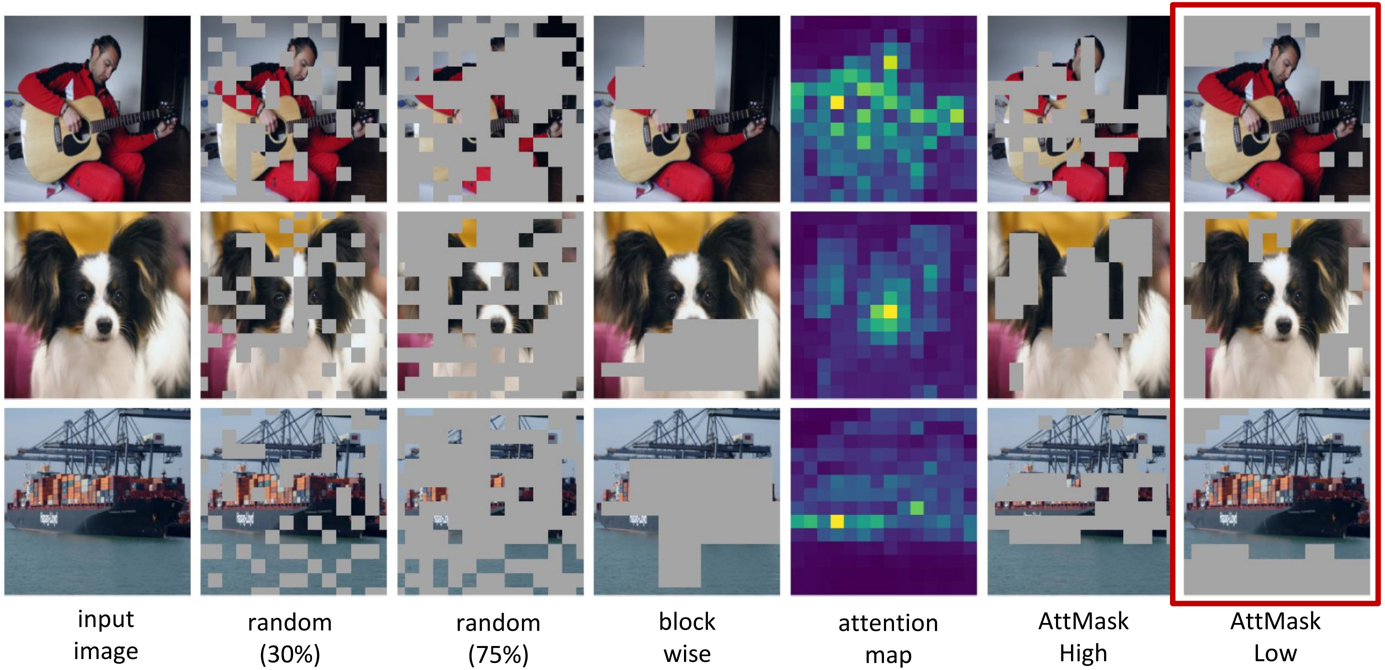
# Approach: Attention-guided token masking (AttMask)

- Leverage ViT's self-attention to mask tokens



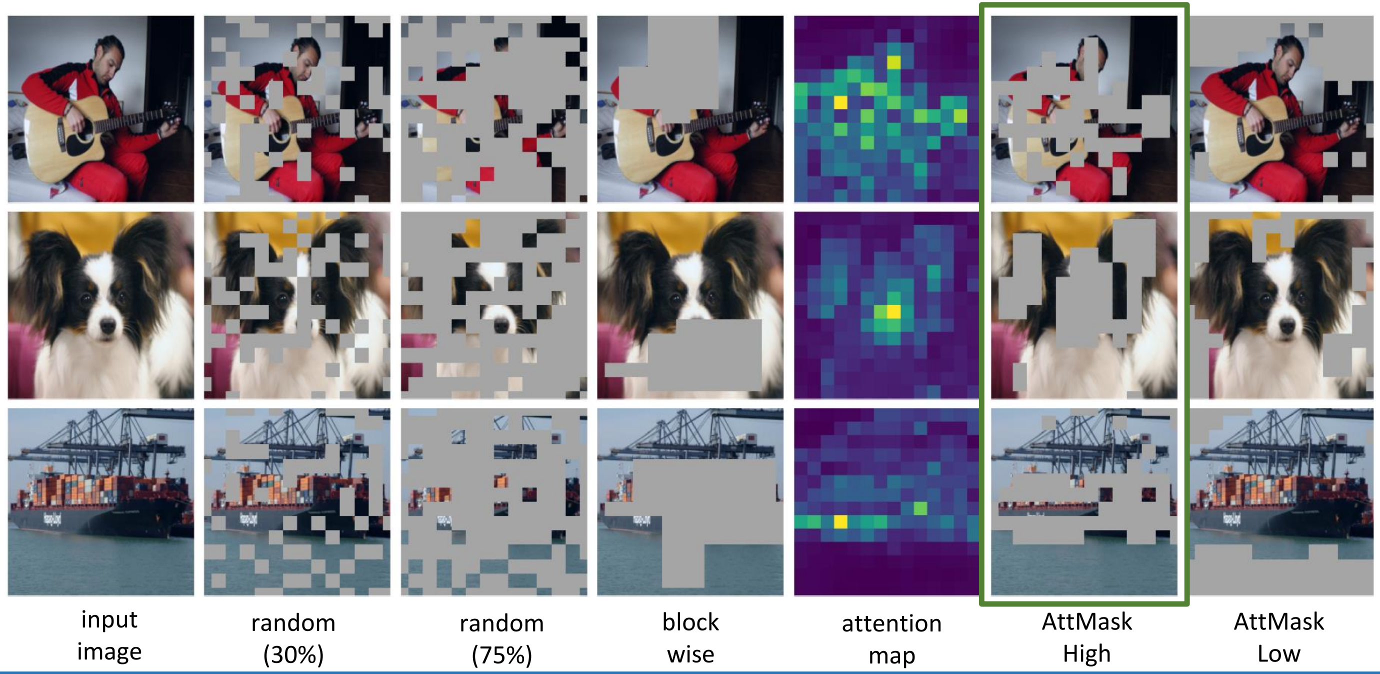
# Approach: Attention-guided token masking (AttMask)

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



# Approach: Attention-guided token masking (AttMask)

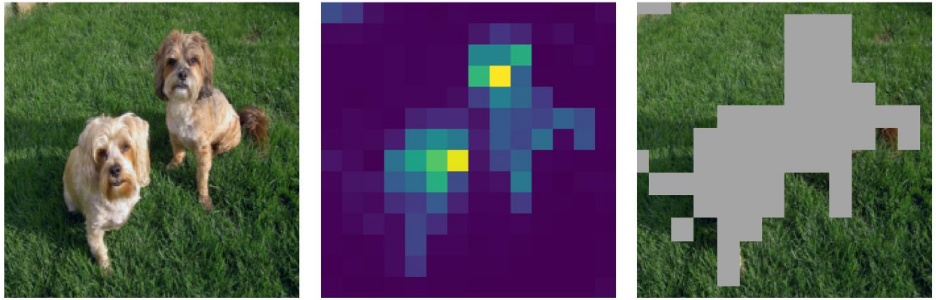
- Leverage ViT's self-attention to mask tokens
  - ✓ **AttMask-High**: masks highly-attended tokens (essentially foreground) →very challenging reconstruction task → boosts performance



# 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!

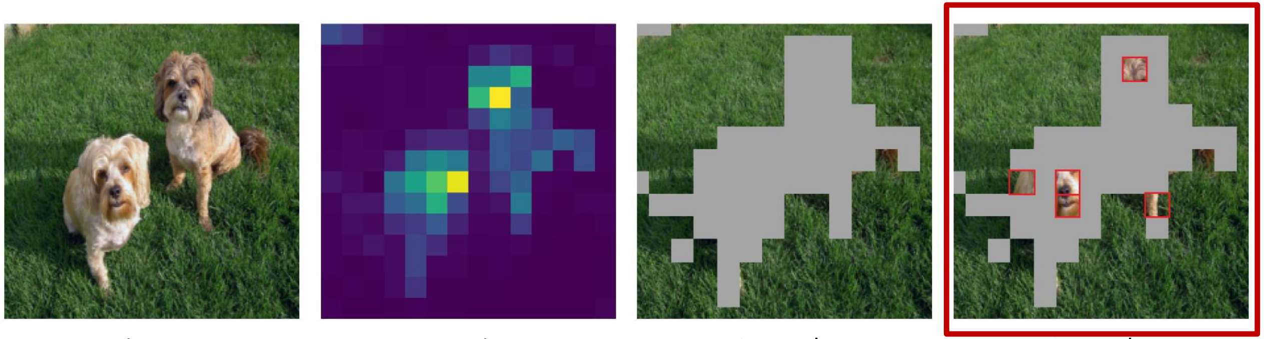


input image | attention map | AttMask High

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

# Approach: Attention-guided token masking (AttMask)

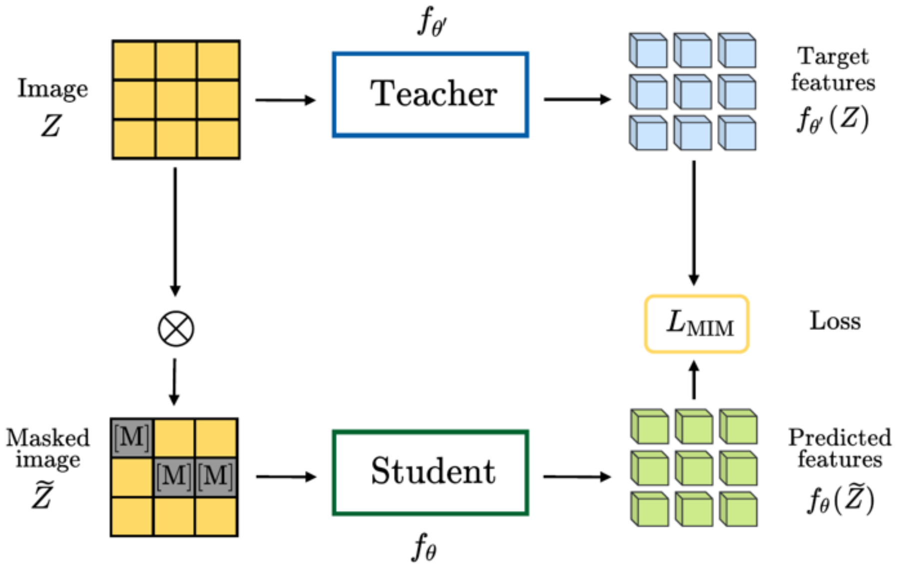
- Leverage ViT's self-attention to mask tokens
  - ✓ **AttMask-High**: masks highly-attended tokens (essentially foreground)
    →very challenging reconstruction task → boosts performance
  - ✓ **AttMask-Hint**: masks highly-attended tokens, but leaves some hints
    →provides hints for the identity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



input image | attention map | AttMask High | AttMask Hint

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

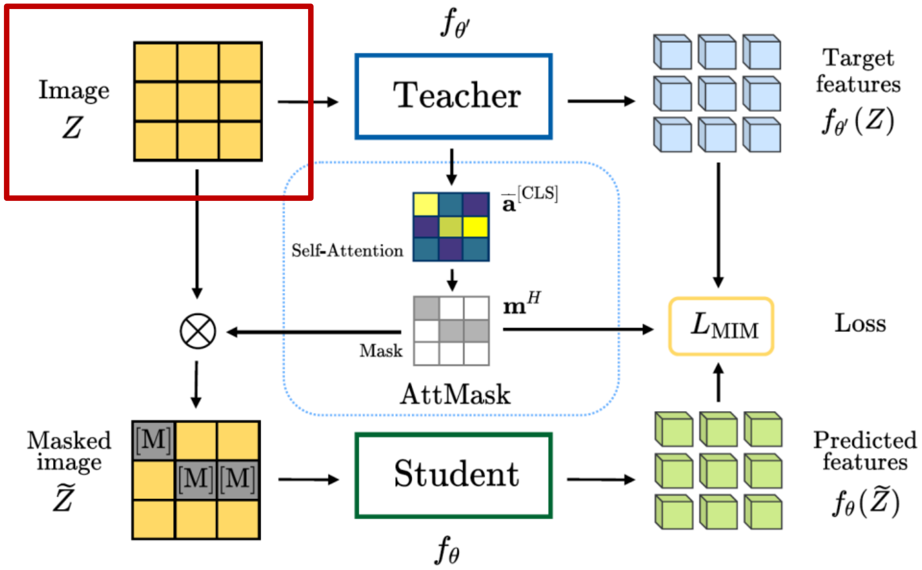
# Incorporating AttMask into distillation-based methods




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

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

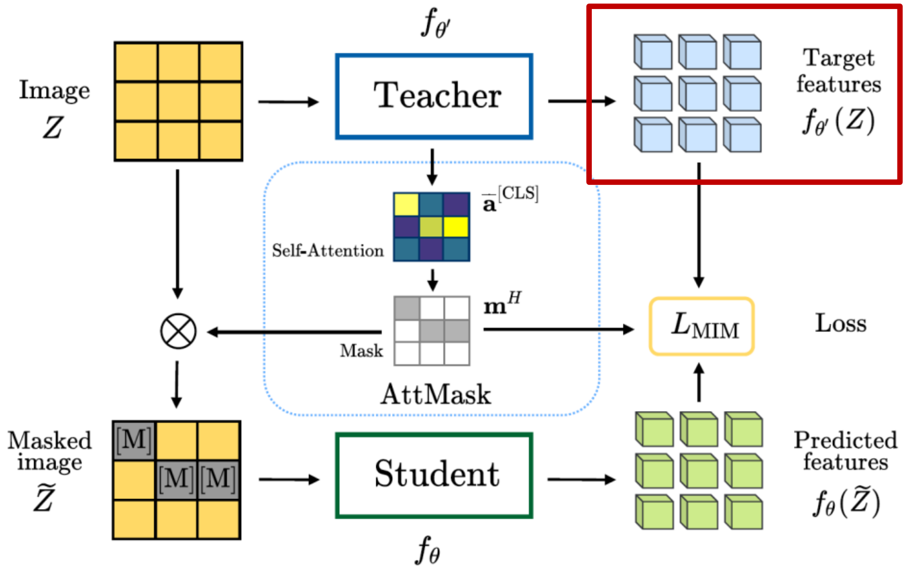
# Incorporating AttMask into distillation-based methods




- 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

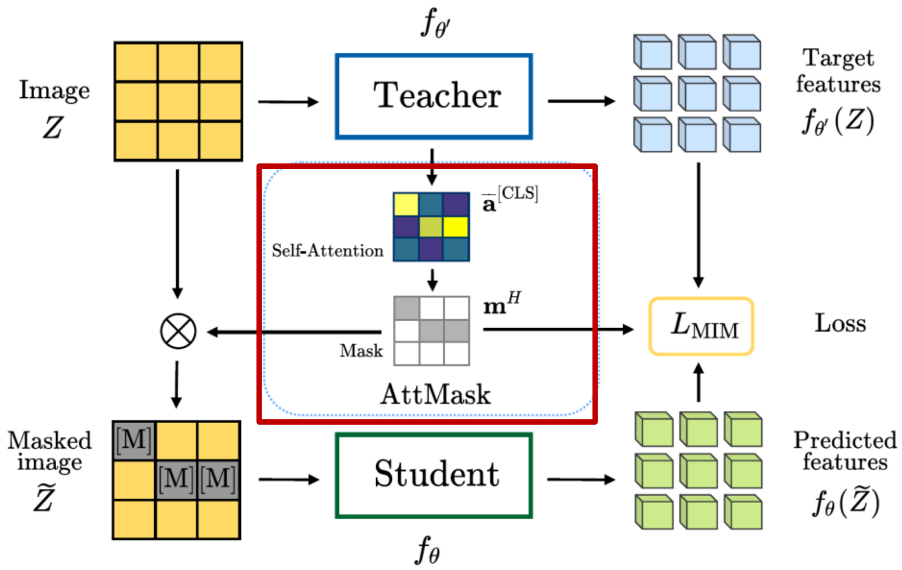
# Incorporating AttMask into distillation-based methods




- 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

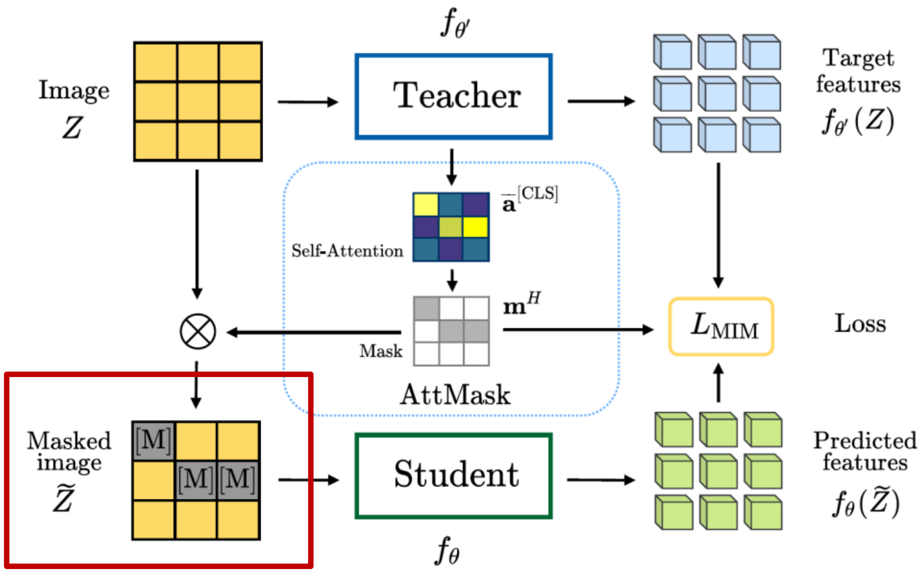
# Incorporating AttMask into distillation-based methods




- 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

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

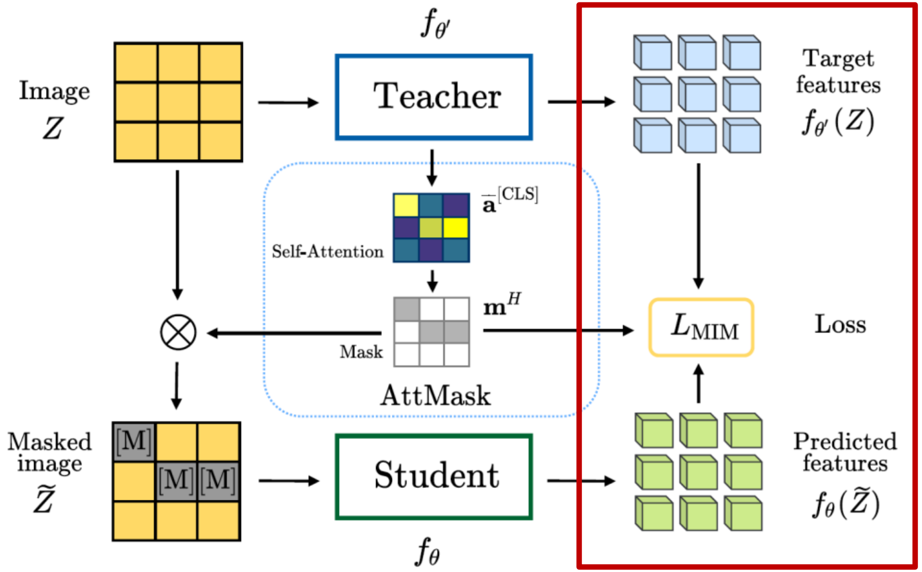
# Incorporating AttMask into distillation-based methods




- 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

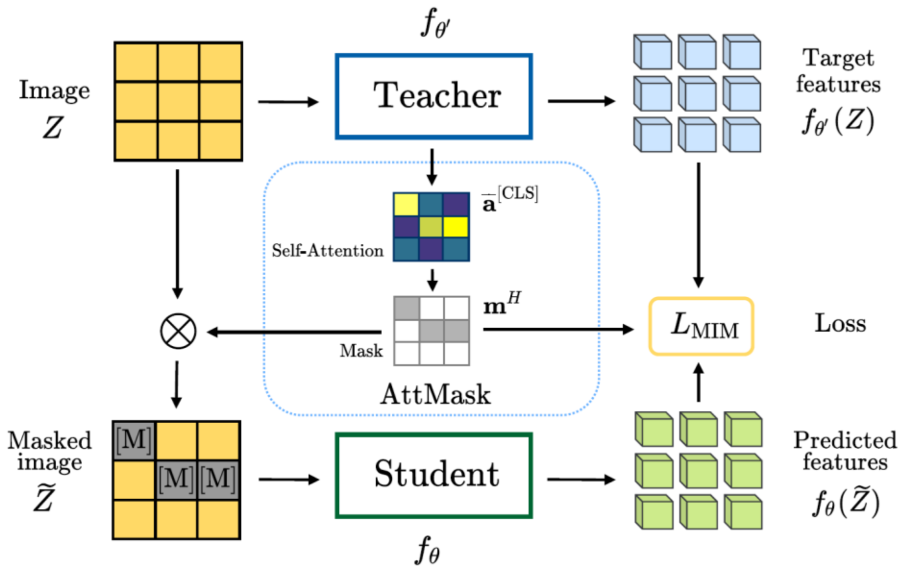
# Incorporating AttMask into distillation-based methods




- 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

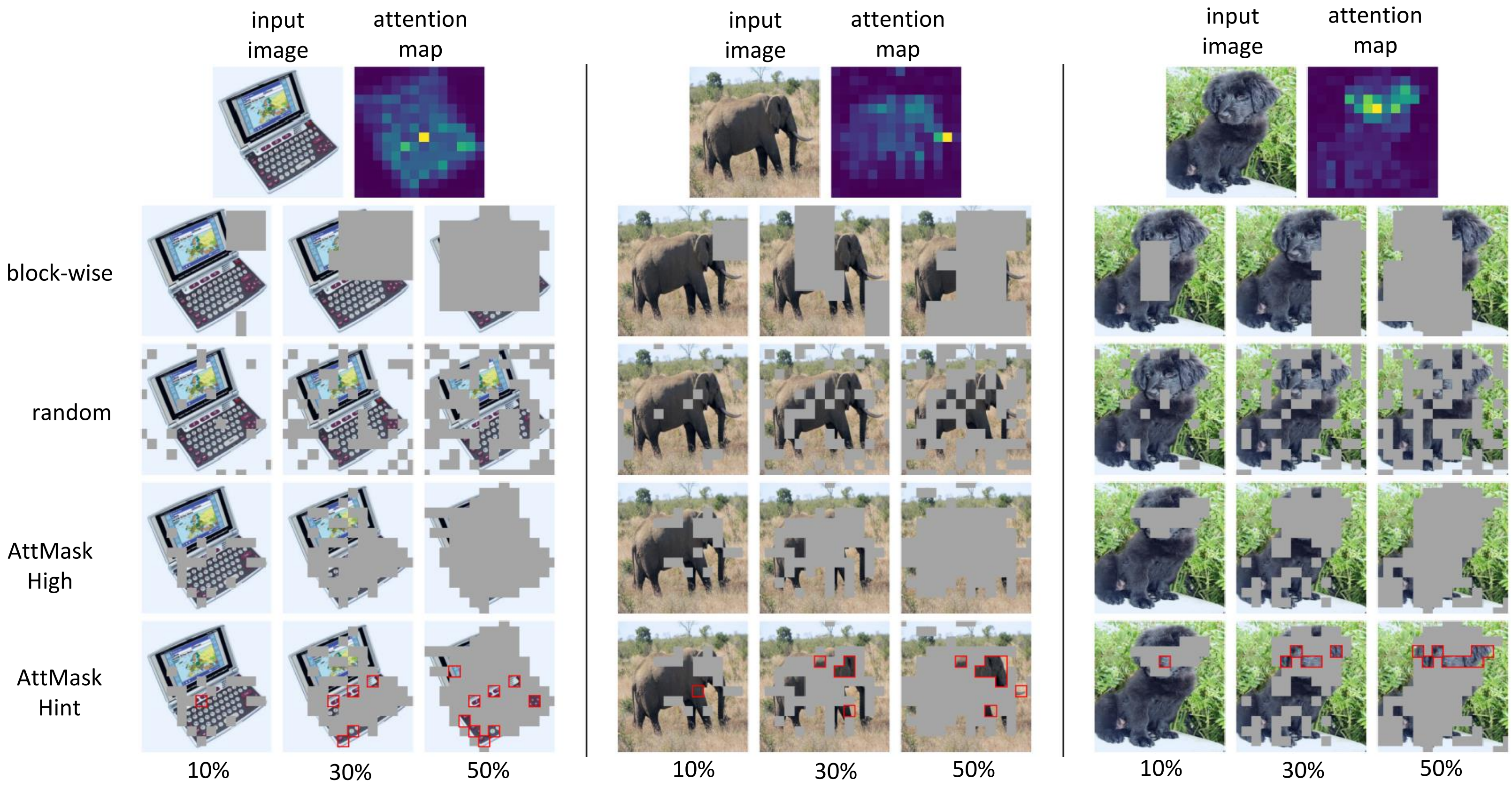
# Incorporating AttMask into distillation-based methods




- 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




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

# Evaluating token masking strategies (20% of ImageNet-1k)

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

| 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)

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

| 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)

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

| 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

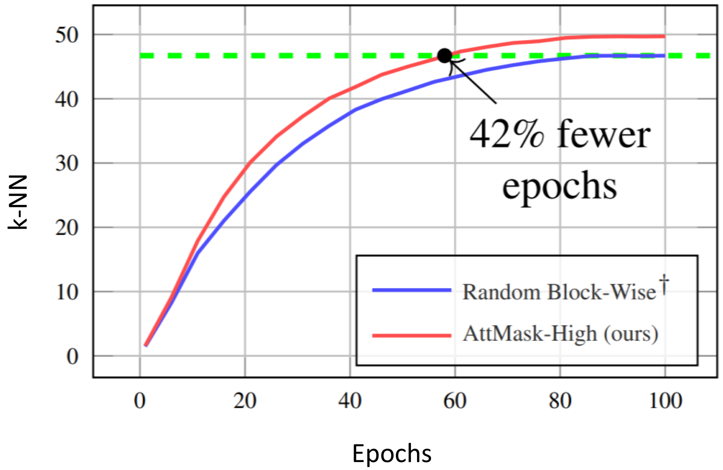
Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022

# Evaluating token masking strategies (20% of ImageNet-1k)

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

| 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
- ✓ 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)

†: 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

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 | 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** |

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

†: 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 | 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** |

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)
- ✓ Evaluating using only 1, 5, 10 or 20 samples per class for the k-NN classifier (more than +3% on low shot k-NN)

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

| Method | COCO | | ADE20K | $\mathcal{R}$Oxford | | $\mathcal{R}$Paris | | DAVIS 2017 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $AP^b$ | $AP^m$ | mIoU | Medium | Hard | Medium | Hard | $(\mathcal{J}\&\mathcal{F})_m$ | $\mathcal{J}_m$ | $\mathcal{F}_m$ |
| 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** |

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

# Property: High-quality features

| METHOD | COCO | | ADE20K | $\mathcal{R}$OXFORD | | $\mathcal{R}$PARIS | | DAVIS 2017 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $AP^b$ | $AP^m$ | mIoU | MEDIUM | HARD | MEDIUM | HARD | $(\mathcal{J}\&\mathcal{F})_m$ | $\mathcal{J}_m$ | $\mathcal{F}_m$ |
| 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** |

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

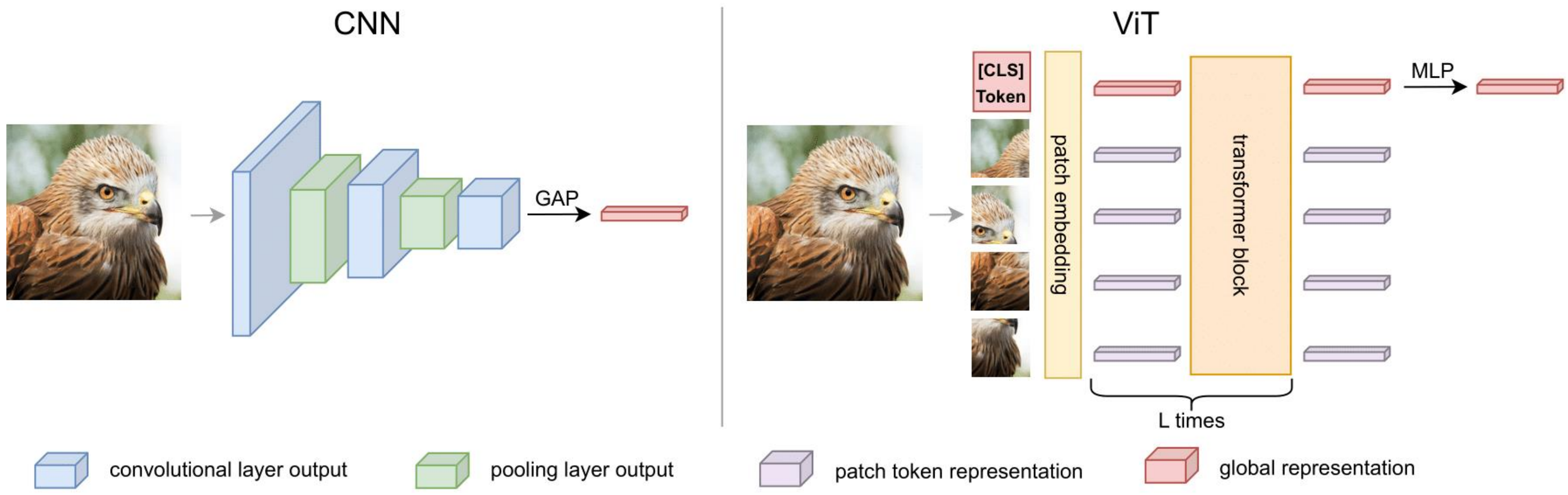
AttMask:

- ✓ 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.

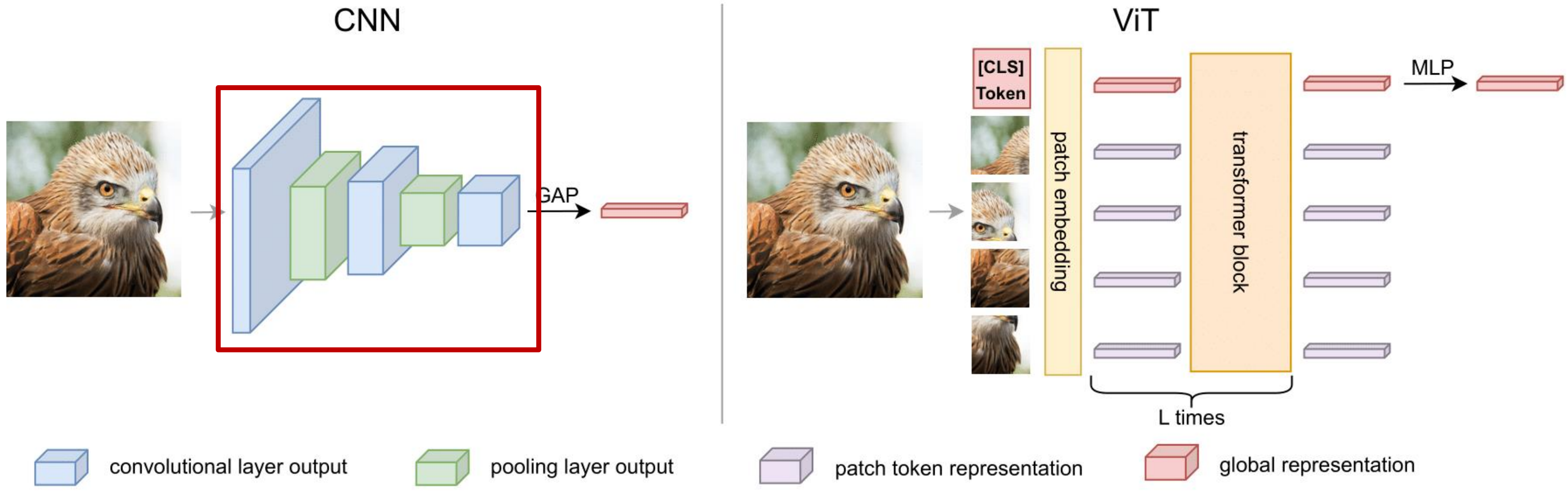
Kakogeorgiou et al., What to Hide from Your Students: Attention-Guided Masked Image Modeling, ECCV 2022

# Leveraging Attention in Pooling

# CNNs vs. ViTs

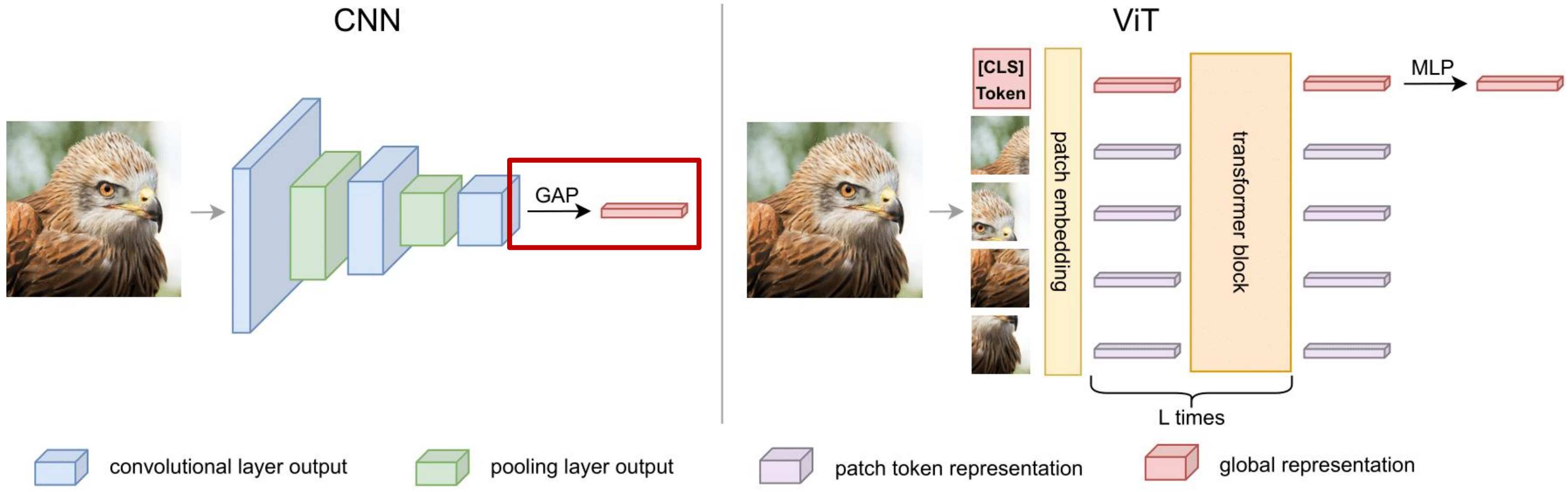

CNN
GAP
ViT
[CLS] Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

# CNNs vs. ViTs

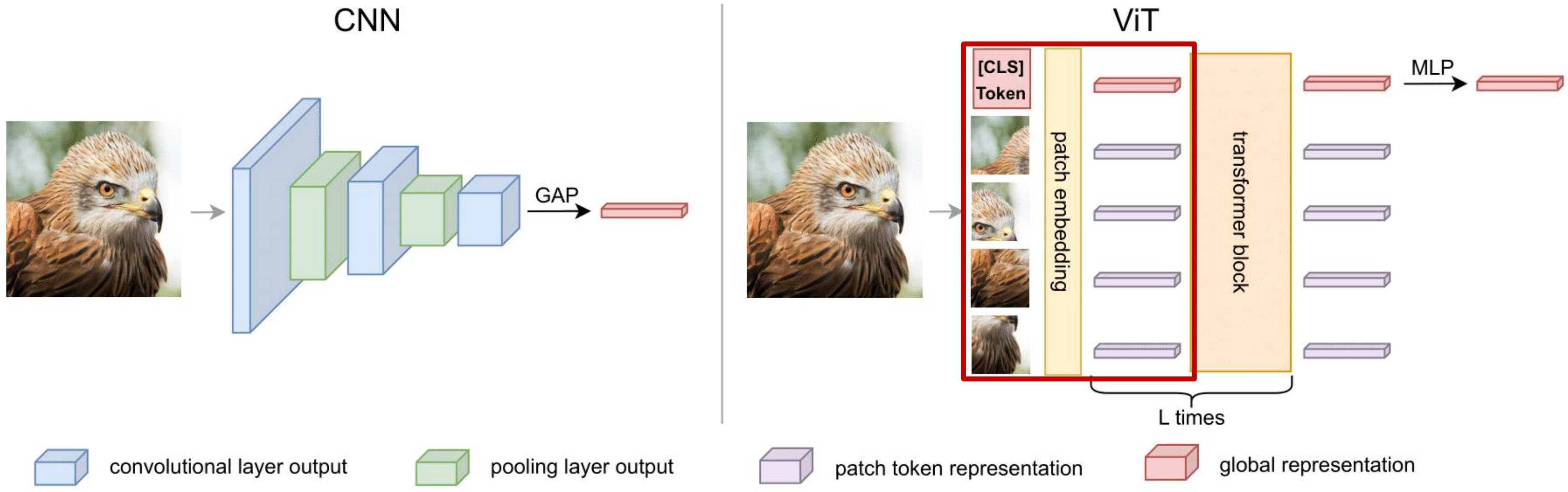


CNN
GAP
ViT
[CLS] Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

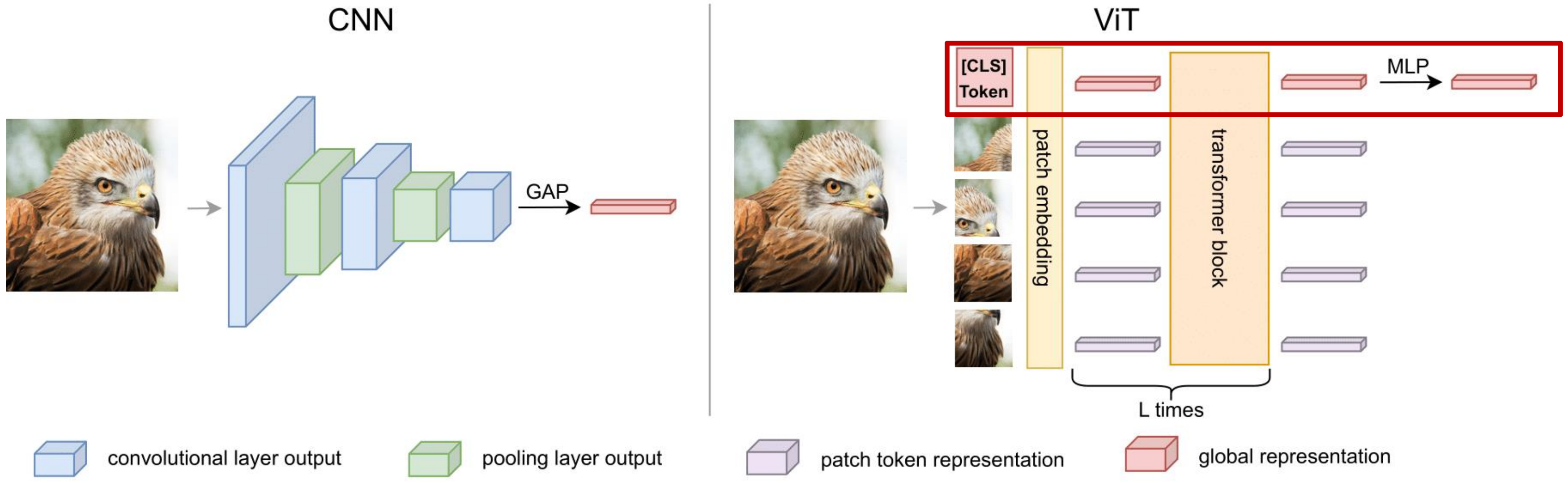
# CNNs vs. ViTs


CNN
GAP
ViT
[CLS]
Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

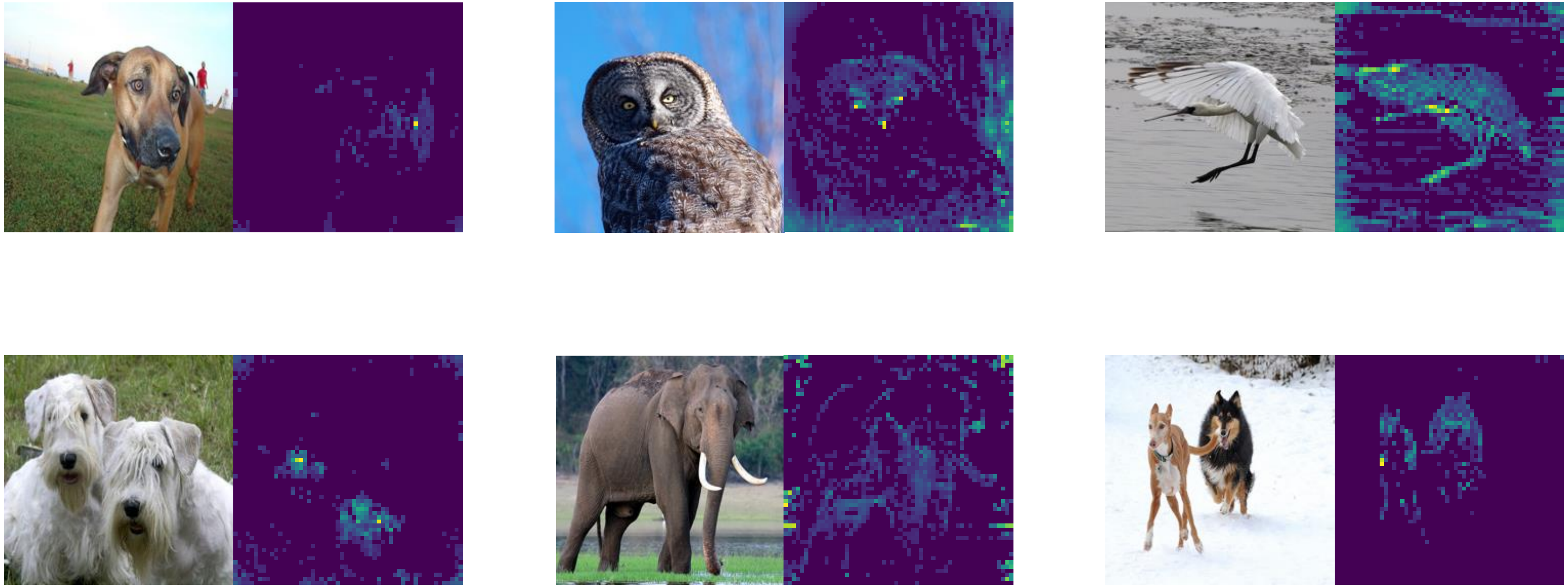
# CNNs vs. ViTs


CNN
GAP
ViT
[CLS] Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

# CNNs vs. ViTs


CNN
GAP
ViT
[CLS] Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

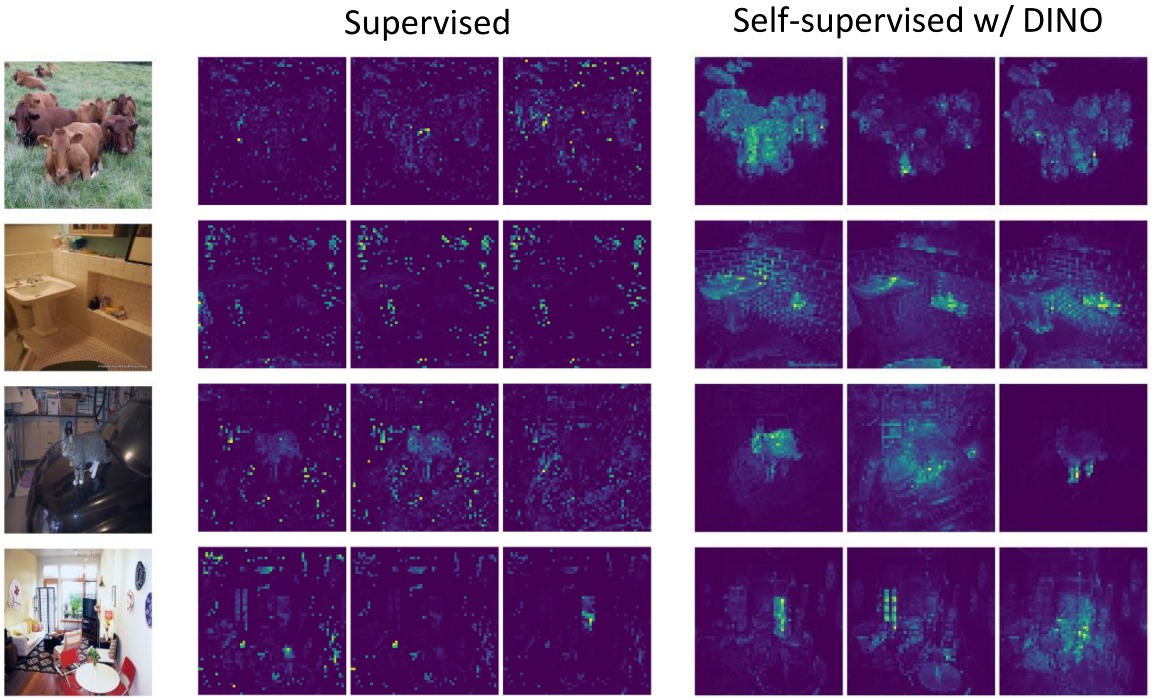
# Supervised ViTs: low-quality attention



ViT-S on Imagenet-1k; mean attention map of the [CLS]; final block

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

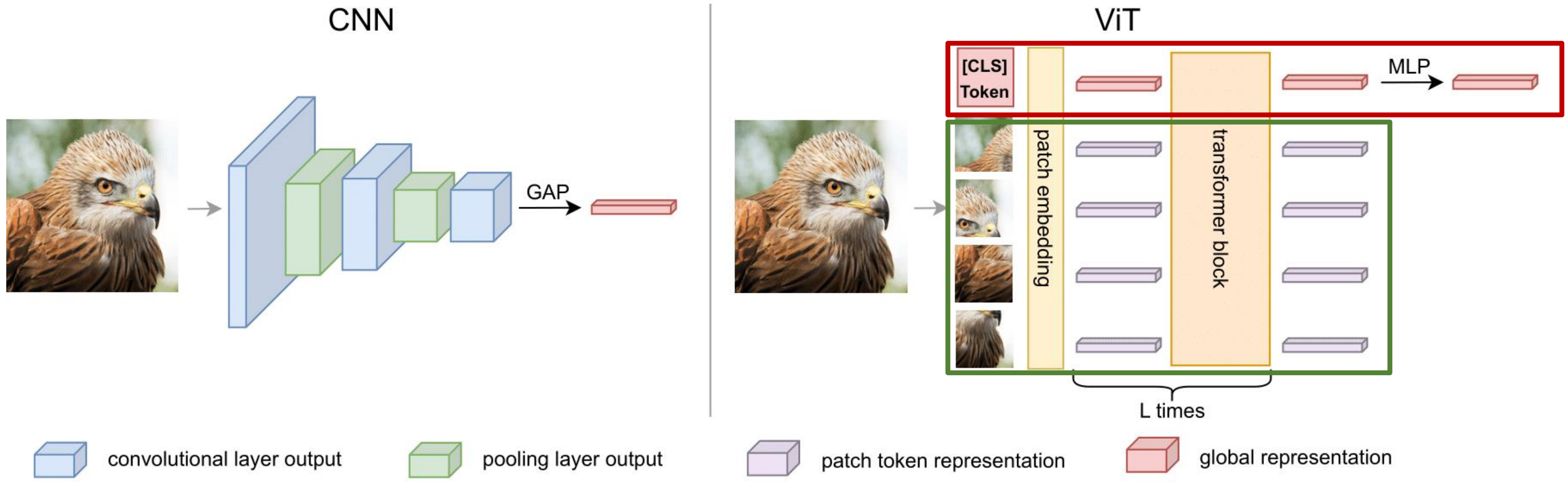
# Is supervision the problem?



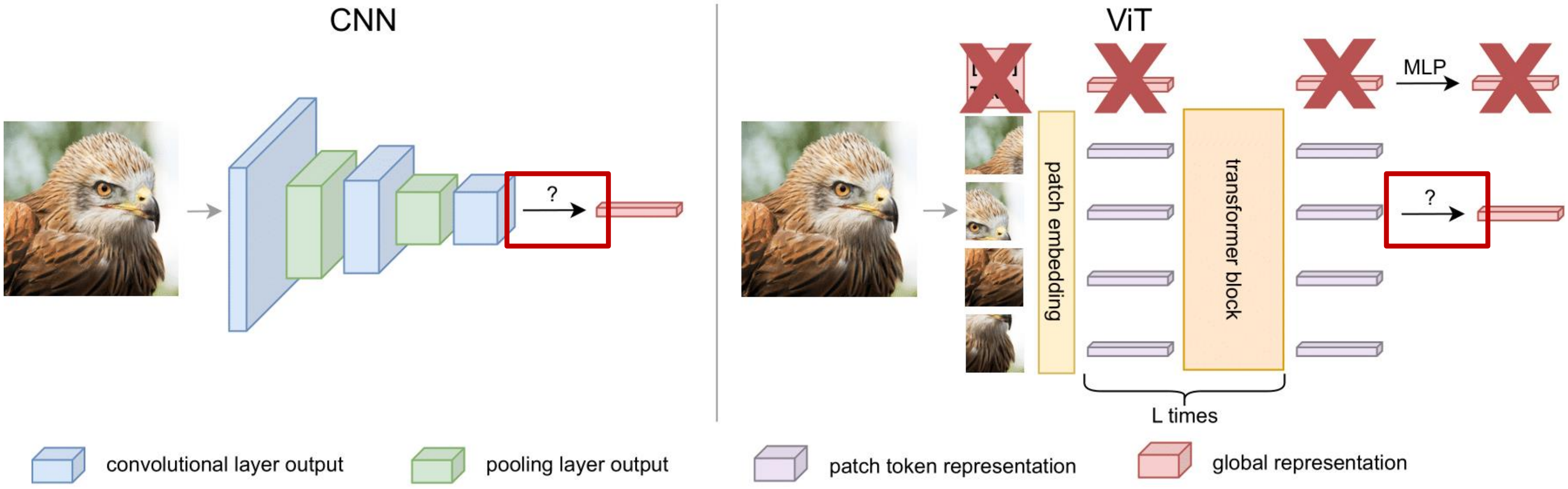

ViT-S on Imagenet-1k; images from COCO val set;
attention maps of the [CLS] for 3 different heads; final block

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

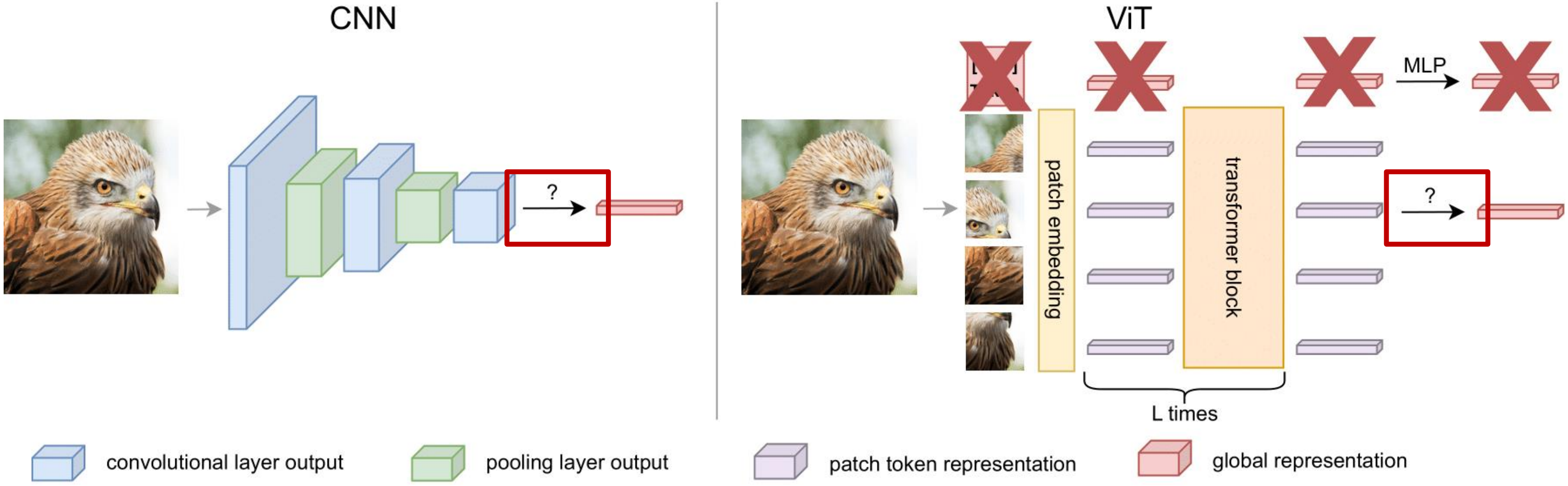
# CNNs vs. ViTs


CNN
GAP
ViT
[CLS] Token
patch embedding
transformer block
MLP
L times
convolutional layer output
pooling layer output
patch token representation
global representation

# “Universal” Pooling

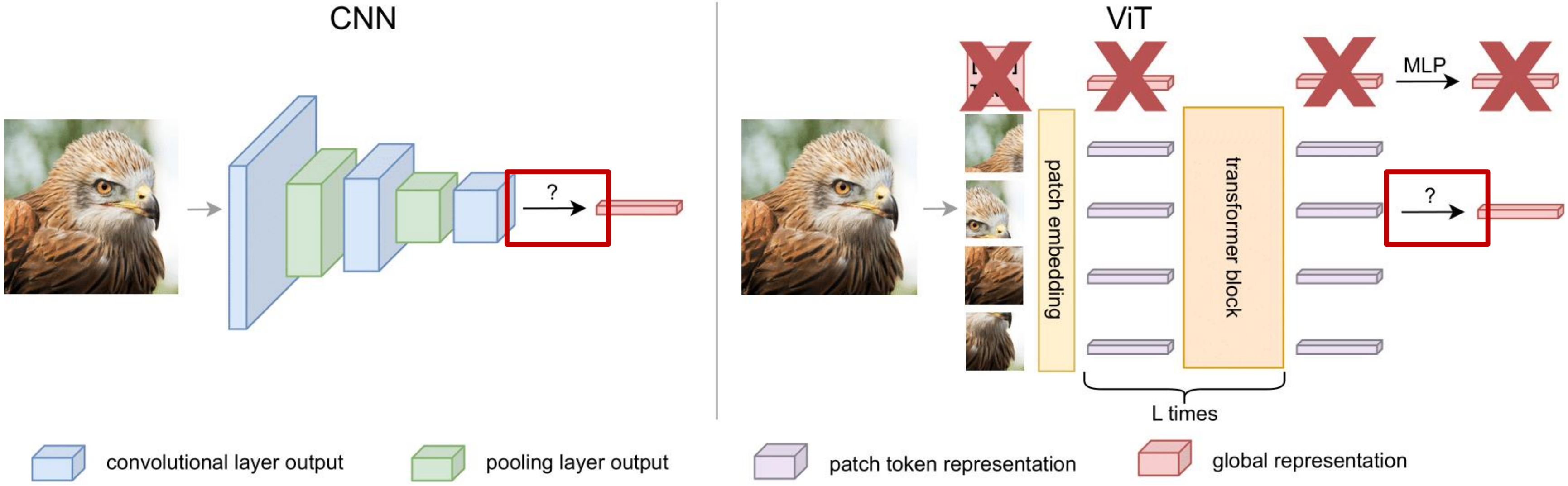


# Focus



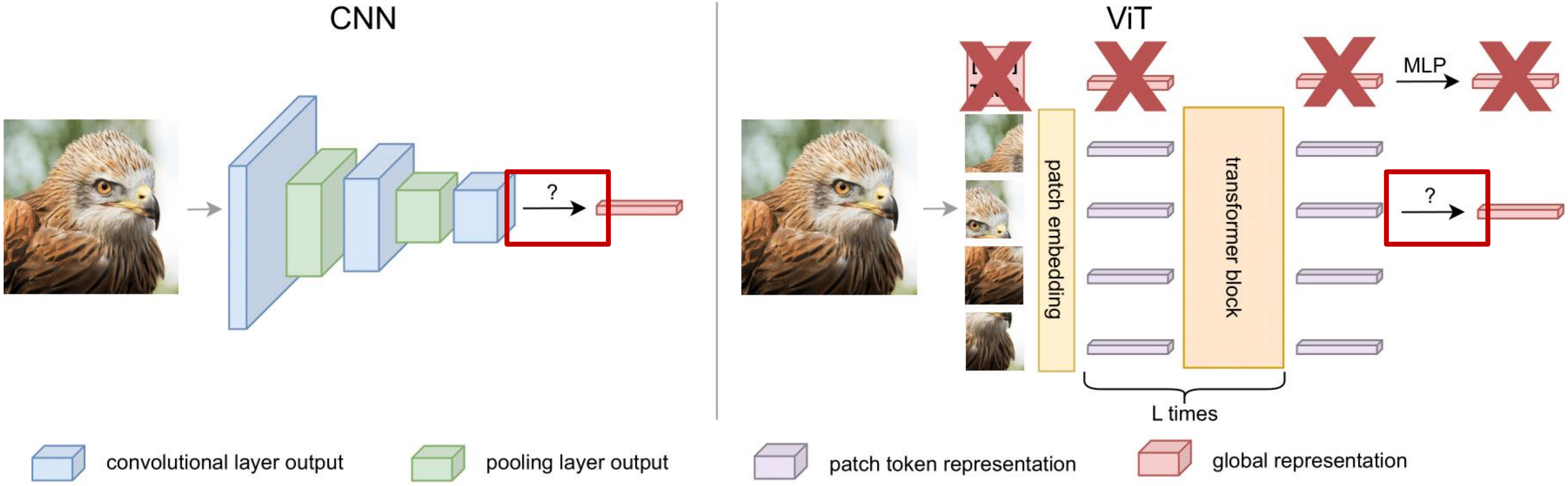

- Pooling at the very last step of both network types improving over default?

# Focus




- Pooling at the very last step of both network types improving over default?
- Pooling for high-quality spatial attention?

# Focus




- Pooling at the very last step of both network types improving over default?
- Pooling for high-quality spatial attention?
- Validity in both supervised and self-supervised settings?

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | |

used in category-level tasks

key mapping

pooling function

# pooled vectors

attention map

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | k | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

used in category-level tasks

# pooled vectors

key mapping

attention map

pooling function

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Generic Pooling Framework

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| # | METHOD | CAT | ITER | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $\boxed{f(x)}$ | $\phi_X(X)$ | $\phi_U(Z)$ | ACC (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

used in category-level tasks

key mapping

pooling function

# pooled vectors

attention map

output mapping

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Formulate methods as instantiations

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

simple, k=1, non-attention

| # | METHOD | CAT | ITER | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x},\mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | ACC (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GAP | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_{-1}(x)$ | | $Z$ | |
| | max | | | 1 | | | | | $\mathbf{1}_p$ | $X$ | $f_{-\infty}(x)$ | | $Z$ | |
| | GeM | | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_\alpha(x)$ | | $Z$ | |
| | LSE | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $e^{rx}$ | | $Z$ | |
| | HOW | | | 1 | | | | | $\operatorname{diag}(X^\top X)$ | $\mathrm{FC}(\mathrm{avg}_3(X))$ | $f_{-1}(x)$ | | $Z$ | |
| 2 | OTK | ✓ | | $k$ | $U$ | $U$ | $X$ | $-\|\mathbf{x}-\mathbf{y}\|^2$ | $\mathrm{SINKHORN}(e^{S/\epsilon})$ | $\psi(X)$ | $f_{-1}(x)$ | | $Z$ | |
| | $k$-means | | ✓ | $k$ | random | $U$ | $X$ | $-\|\mathbf{x}-\mathbf{y}\|^2$ | $\eta_2(\arg\max_1(S))$ | $X$ | $f_{-1}(x)$ | $X$ | $Z$ | |
| | Slot* | ✓ | ✓ | $k$ | $U$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S/\sqrt{d})$ | $W_V X$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(\mathrm{GRU}(Z))$ | |
| 3 | SE | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | | | | $\operatorname{diag}(\mathbf{q})X$ | | $V$ | | |
| | CBAM* | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | $X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma(\mathrm{conv}_7(S))$ | $\operatorname{diag}(\mathbf{q})X$ | | $V\operatorname{diag}(\mathbf{a})$ | | |
| 4 | ViT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $\mathrm{MLP}(\mathrm{MSA}(X))$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |
| | CaiT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |

used in category-level tasks

key mapping

pooling function

attention map

output mapping

# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Formulate methods as instantiations

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

simple, k=1, non-attention

k>1

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x}, \mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GAP | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_{-1}(x)$ | | $Z$ | |
| | max | | | 1 | | | | | $\mathbf{1}_p$ | $X$ | $f_{-\infty}(x)$ | | $Z$ | |
| | GeM | | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_\alpha(x)$ | | $Z$ | |
| | LSE | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $e^{rx}$ | | $Z$ | |
| | HOW | | | 1 | | | | | $\operatorname{diag}(X^\top X)$ | $\operatorname{FC}(\operatorname{avg}_3(X))$ | $f_{-1}(x)$ | | $Z$ | |
| 2 | OTK | ✓ | | $k$ | $U$ | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\operatorname{Sinkhorn}(e^{S/\epsilon})$ | $\psi(X)$ | $f_{-1}(x)$ | | $Z$ | |
| | $k$-means | | ✓ | $k$ | random | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\eta_2(\arg\max_1(S))$ | $X$ | $f_{-1}(x)$ | $X$ | $Z$ | |
| | Slot* | ✓ | ✓ | $k$ | $U$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S/\sqrt{d})$ | $W_V X$ | $f_{-1}(x)$ | $X$ | $\operatorname{MLP}(\operatorname{GRU}(Z))$ | |
| 3 | SE | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\operatorname{MLP}(U))$ | | | | $\operatorname{diag}(\mathbf{q})X$ | | $V$ | | |
| | CBAM* | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\operatorname{MLP}(U))$ | $X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma(\operatorname{conv}_7(S))$ | $\operatorname{diag}(\mathbf{q})X$ | | $V\operatorname{diag}(\mathbf{a})$ | | |
| 4 | ViT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $\operatorname{MLP}(\operatorname{MSA}(X))$ | $\operatorname{MLP}(g_m^{-1}(Z))$ | |
| | CaiT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $X$ | $\operatorname{MLP}(g_m^{-1}(Z))$ | |

used in category-level tasks

key mapping

pooling function

attention map

output mapping

# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Formulate methods as instantiations

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

simple, k=1, non-attention

k>1

modules within arch.

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x},\mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GAP | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_{-1}(x)$ | | $Z$ | |
| | max | | | 1 | | | | | $\mathbf{1}_p$ | $X$ | $f_{-\infty}(x)$ | | $Z$ | |
| | GeM | | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_\alpha(x)$ | | $Z$ | |
| | LSE | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $e^{rx}$ | | $Z$ | |
| | HOW | | | 1 | | | | | $\mathrm{diag}(X^\top X)$ | $\mathrm{FC}(\mathrm{avg}_3(X))$ | $f_{-1}(x)$ | | $Z$ | |
| 2 | OTK | ✓ | | $k$ | $U$ | $U$ | $X$ | $-\|\mathbf{x}-\mathbf{y}\|^2$ | $\mathrm{Sinkhorn}(e^{S/\epsilon})$ | $\psi(X)$ | $f_{-1}(x)$ | | $Z$ | |
| | $k$-means | | ✓ | $k$ | random | $U$ | $X$ | $-\|\mathbf{x}-\mathbf{y}\|^2$ | $\eta_2(\arg\max_1(S))$ | $X$ | $f_{-1}(x)$ | $X$ | $Z$ | |
| | Slot* | ✓ | ✓ | $k$ | $U$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S/\sqrt{d})$ | $W_V X$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(\mathrm{GRU}(Z))$ | |
| 3 | SE | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | | | | $\mathrm{diag}(\mathbf{q})X$ | | $V$ | | |
| | CBAM* | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | $X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma(\mathrm{conv}_7(S))$ | $\mathrm{diag}(\mathbf{q})X$ | | $V\,\mathrm{diag}(\mathbf{a})$ | | |
| 4 | ViT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $\mathrm{MLP}(\mathrm{MSA}(X))$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |
| | CaiT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |

used in category-level tasks

key mapping

pooling function

attention map

output mapping

# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Formulate methods as instantiations

iterative
query mapping
value mapping
output mapping
init. pooled vectors
similarity function
classification accuracy on ImageNet-1k

| | # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x},\mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| simple, k=1, non-attention | 1 | GAP | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_{-1}(x)$ | | $Z$ | |
| | | max | | | 1 | | | | | $\mathbf{1}_p$ | $X$ | $f_{-\infty}(x)$ | | $Z$ | |
| | | GeM | | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_\alpha(x)$ | | $Z$ | |
| | | LSE | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $e^{rx}$ | | $Z$ | |
| | | HOW | | | 1 | | | | | $\operatorname{diag}(X^\top X)$ | $\mathrm{FC}(\mathrm{avg}_3(X))$ | $f_{-1}(x)$ | | $Z$ | |
| k>1 | 2 | OTK | ✓ | | $k$ | $U$ | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\mathrm{Sinkhorn}(e^{S/\epsilon})$ | $\psi(X)$ | $f_{-1}(x)$ | | $Z$ | |
| | | $k$-means | | ✓ | $k$ | random | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\eta_2(\arg\max_1(S))$ | $X$ | $f_{-1}(x)$ | $X$ | $Z$ | |
| | | Slot* | ✓ | ✓ | $k$ | $U$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S/\sqrt{d})$ | $W_V X$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(\mathrm{GRU}(Z))$ | |
| modules within arch. | 3 | SE | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | | | | $\operatorname{diag}(\mathbf{q})X$ | | $V$ | | |
| | | CBAM* | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | $X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma(\mathrm{conv}_7(S))$ | $\operatorname{diag}(\mathbf{q})X$ | | $V\operatorname{diag}(\mathbf{a})$ | | |
| vision transformers | 4 | ViT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $\mathrm{MLP}(\mathrm{MSA}(X))$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |
| | | CaiT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |

used in category-level tasks
key mapping
pooling function
attention map
output mapping
# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Discuss and derive

iterative

query mapping

value mapping

output mapping

init. pooled vectors

similarity function

classification accuracy on ImageNet-1k

| # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x},\mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GAP | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_{-1}(x)$ | | $Z$ | |
| | max | | | 1 | | | | | $\mathbf{1}_p$ | $X$ | $f_{-\infty}(x)$ | | $Z$ | |
| | GeM | | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_\alpha(x)$ | | $Z$ | |
| | LSE | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $e^{rx}$ | | $Z$ | |
| | HOW | | | 1 | | | | | $\operatorname{diag}(X^\top X)$ | $\mathrm{FC}(\mathrm{avg}_3(X))$ | $f_{-1}(x)$ | | $Z$ | |
| 2 | OTK | ✓ | | $k$ | $U$ | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\mathrm{Sinkhorn}(e^{S/\epsilon})$ | $\psi(X)$ | $f_{-1}(x)$ | | $Z$ | |
| | $k$-means | | ✓ | $k$ | random | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\eta_2(\arg\max_1(S))$ | $X$ | $f_{-1}(x)$ | $X$ | $Z$ | |
| | Slot* | ✓ | ✓ | $k$ | $U$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S/\sqrt{d})$ | $W_V X$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(\mathrm{GRU}(Z))$ | |
| 3 | SE | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | | | | $\operatorname{diag}(\mathbf{q})X$ | | $V$ | | |
| | CBAM* | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | $X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma(\mathrm{conv}_7(S))$ | $\operatorname{diag}(\mathbf{q})X$ | | $V\operatorname{diag}(\mathbf{a})$ | | |
| 4 | ViT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $\mathrm{MLP}(\mathrm{MSA}(X))$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |
| | CaiT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | |
| 5 | SimPool | ✓ | | 1 | $\pi_A(X)$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma_2(S/\sqrt{d})$ | $X - \min X$ | $f_\alpha(x)$ | | $Z$ | |

simple, k=1, non-attention

k>1

modules within arch.

vision transformers

used in category-level tasks

key mapping

pooling function

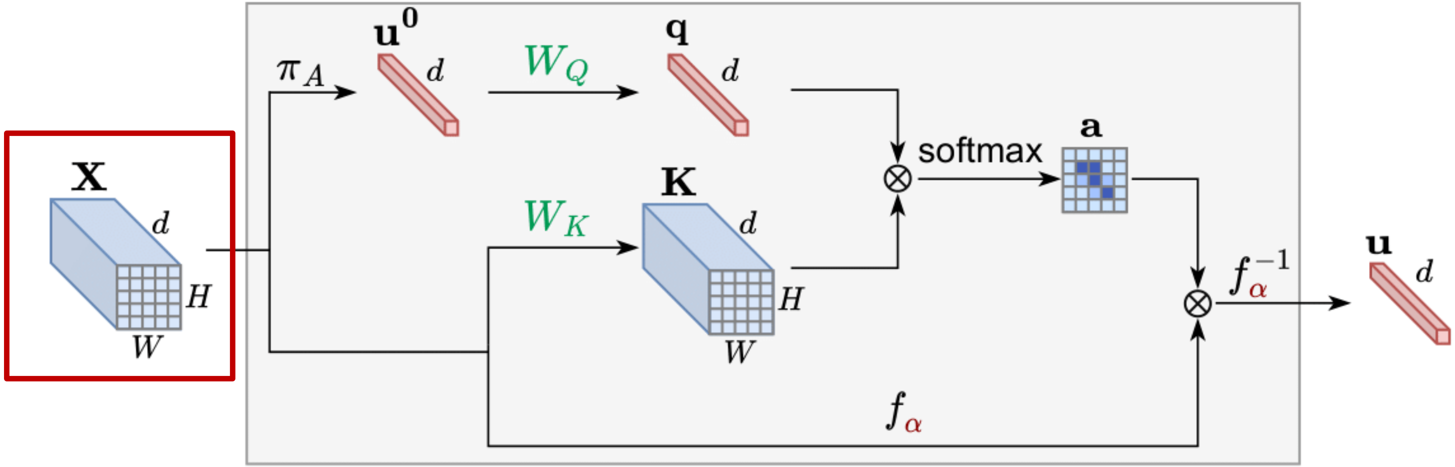
attention map

output mapping

# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool

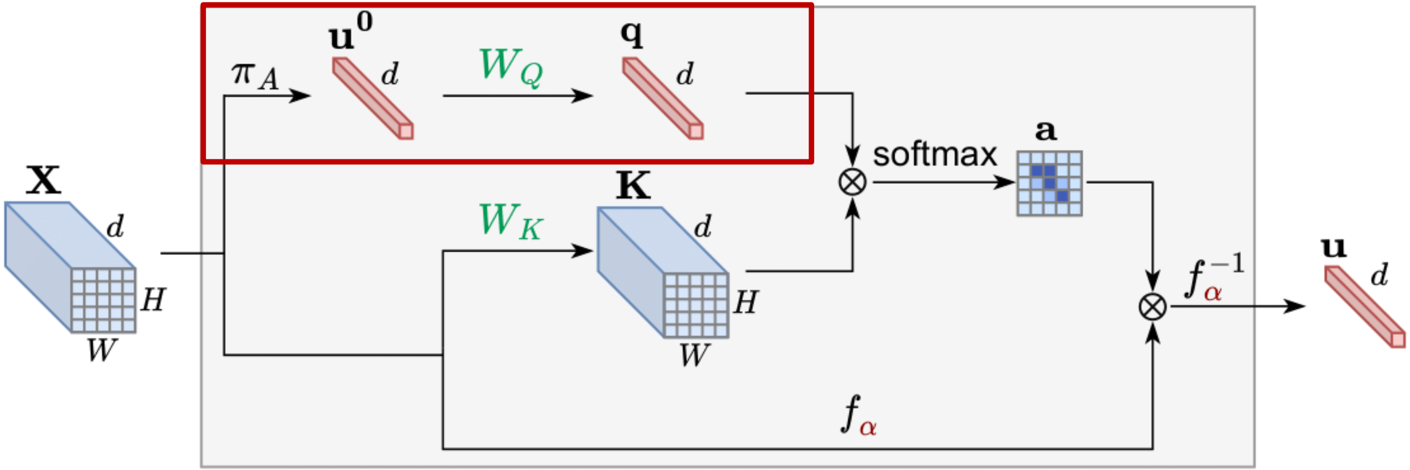



- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2 \left( K^\top \mathbf{q} / \sqrt{d} \right)$.
- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool



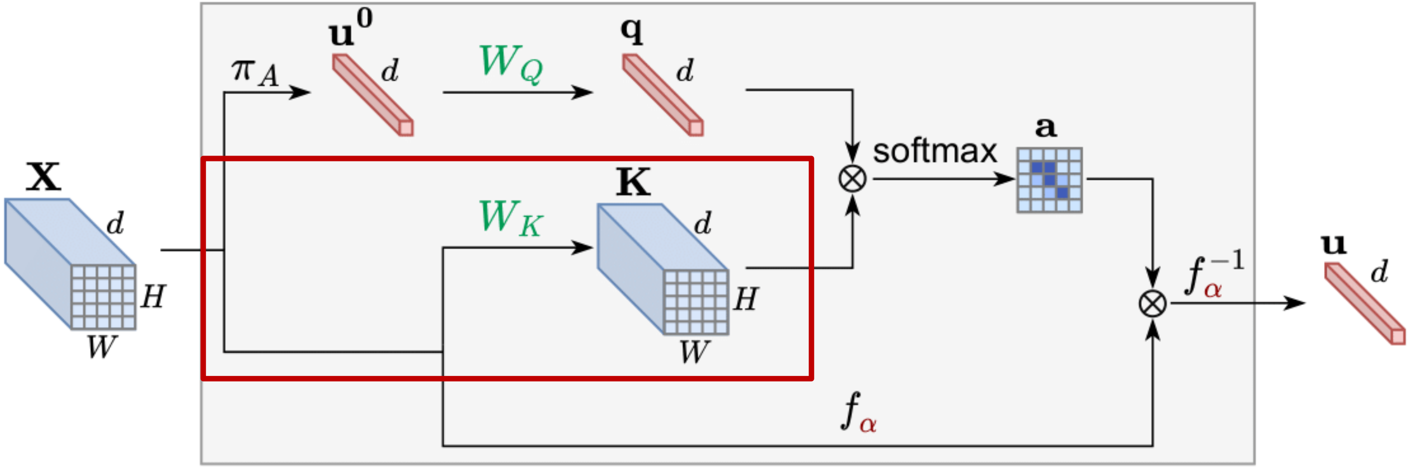

- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2 \left( K^\top \mathbf{q} / \sqrt{d} \right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool



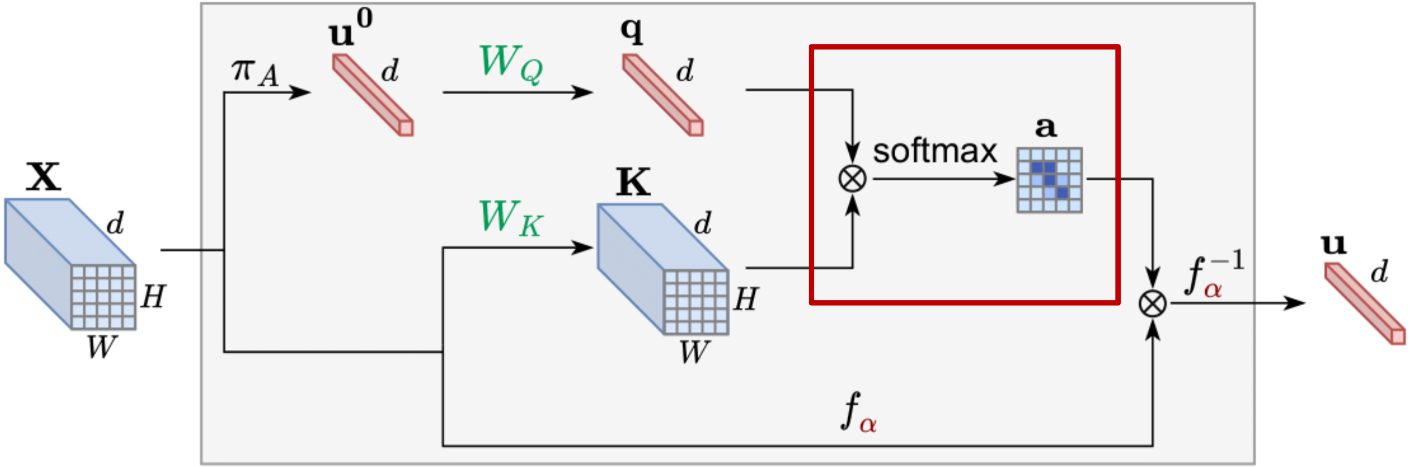

- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2 \left( K^\top \mathbf{q} / \sqrt{d} \right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool

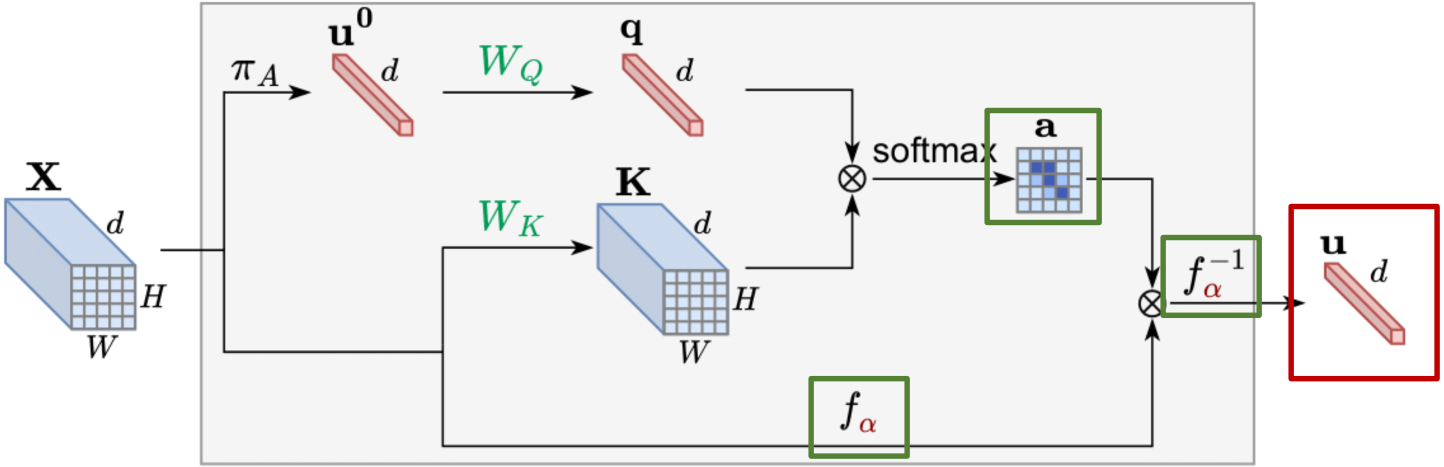



- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top \mathbf{q}/\sqrt{d}\right)$.
- Global representation: $\mathbf{u} = \pi_{\mathrm{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool



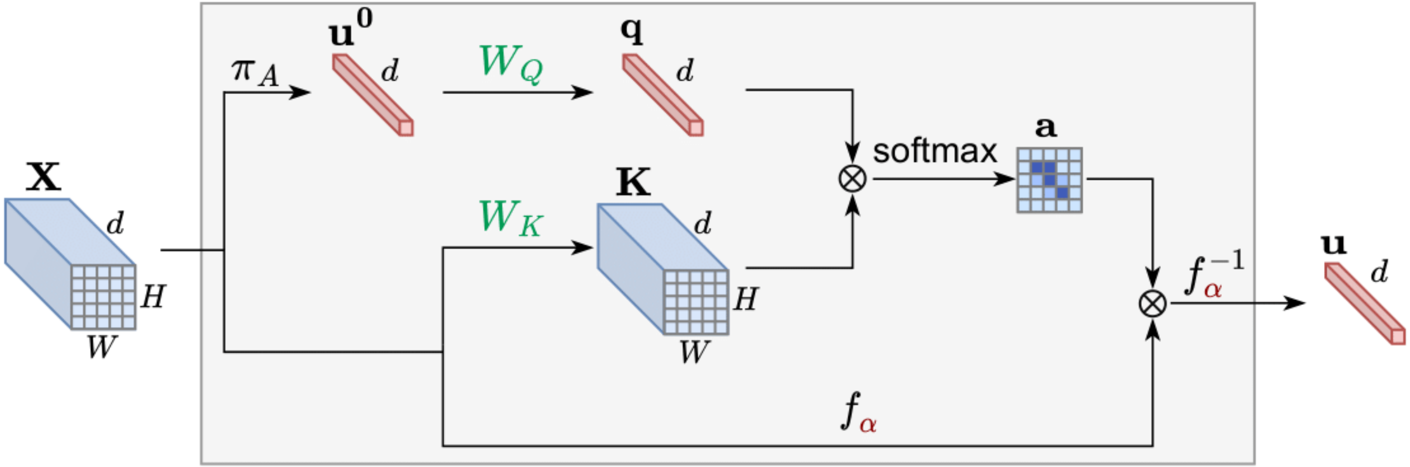

- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top \mathbf{q}/\sqrt{d}\right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool



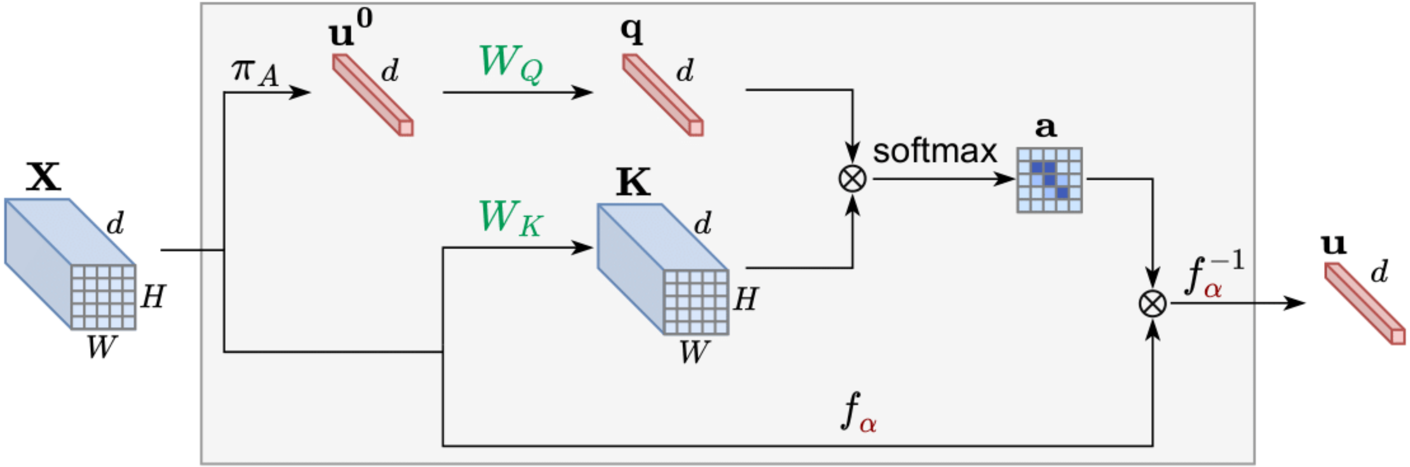

- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2\left(K^\top \mathbf{q}/\sqrt{d}\right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$\boxed{f_\alpha(x) := \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# SimPool




- Initial representation: $\mathbf{u}^0 = \pi_A$ by GAP.
- $\mathbf{u}^0$ ($\mathbf{X}$) mapped by $W_Q$ ($W_K$) to form $\mathbf{q}$ ($\mathbf{K}$).
- Attention map: $\mathbf{a} = \boldsymbol{\sigma}_2 \left( K^\top \mathbf{q} / \sqrt{d} \right)$.

- Global representation: $\mathbf{u} = \pi_{\text{SP}}(X) := f_\alpha^{-1}(f_\alpha(V)\mathbf{a})$, where:

$$f_\alpha(x) := \begin{cases} x^{\boxed{\frac{1-\alpha}{2}}}, & \text{if } \alpha \neq 1, \\ \ln x, & \text{if } \alpha = 1. \end{cases}$$

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Benchmark

iterative

query mapping

init. pooled vectors

similarity function

value mapping

output mapping

classification accuracy on ImageNet-1k

| | # | Method | Cat | Iter | $k$ | $U^0$ | $\phi_Q(U)$ | $\phi_K(X)$ | $s(\mathbf{x},\mathbf{y})$ | $A$ | $\phi_V(X)$ | $f(x)$ | $\phi_X(X)$ | $\phi_U(Z)$ | Acc (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| simple, k=1, non-attention | 1 | GAP | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_{-1}(x)$ | | $Z$ | 55.0 |
| | | max | | | 1 | | | | | $\mathbf{1}_p$ | $X$ | $f_{-\infty}(x)$ | | $Z$ | 53.9 |
| | | GeM | | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $f_\alpha(x)$ | | $Z$ | 55.9 |
| | | LSE | ✓ | | 1 | | | | | $\mathbf{1}_p/p$ | $X$ | $e^{rx}$ | | $Z$ | 55.3 |
| | | HOW | | | 1 | | | | | $\operatorname{diag}(X^\top X)$ | $\mathrm{FC}(\mathrm{avg}_3(X))$ | $f_{-1}(x)$ | | $Z$ | 54.8 |
| k>1 | 2 | OTK | ✓ | | $k$ | $U$ | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\textsc{Sinkhorn}(e^{S/\epsilon})$ | $\psi(X)$ | $f_{-1}(x)$ | | $Z$ | 55.9 |
| | | $k$-means | | ✓ | $k$ | random | $U$ | $X$ | $-\lVert\mathbf{x}-\mathbf{y}\rVert^2$ | $\eta_2(\arg\max_1(S))$ | $X$ | $f_{-1}(x)$ | $X$ | $Z$ | 55.4 |
| | | Slot* | ✓ | ✓ | $k$ | $U$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\boldsymbol{\sigma}_2(S/\sqrt{d})$ | $W_V X$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(\mathrm{GRU}(Z))$ | 56.7 |
| modules within arch. | 3 | SE | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | | | | $\operatorname{diag}(\mathbf{q})X$ | | $V$ | | 55.7 |
| | | CBAM* | ✓ | | 1 | $\pi_A(X)$ | $\sigma(\mathrm{MLP}(U))$ | $X$ | $\mathbf{x}^\top\mathbf{y}$ | $\sigma(\mathrm{conv}_7(S))$ | $\operatorname{diag}(\mathbf{q})X$ | | $V\operatorname{diag}(\mathbf{a})$ | | 55.6 |
| vision transformers | 4 | ViT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $\mathrm{MLP}(\mathrm{MSA}(X))$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | 56.1 |
| | | CaiT* | ✓ | ✓ | 1 | $U$ | $g_m(W_Q U)$ | $g_m(W_K X)$ | $\mathbf{x}^\top\mathbf{y}$ | $\boldsymbol{\sigma}_2(S_i/\sqrt{d})_{i=1}^m$ | $g_m(W_V X)$ | $f_{-1}(x)$ | $X$ | $\mathrm{MLP}(g_m^{-1}(Z))$ | 56.7 |
| | 5 | SimPool | ✓ | | 1 | $\pi_A(X)$ | $W_Q U$ | $W_K X$ | $\mathbf{x}^\top\mathbf{y}$ | $\boldsymbol{\sigma}_2(S/\sqrt{d})$ | $X - \min X$ | $f_\alpha(x)$ | | $Z$ | 57.1 |

used in category-level tasks

key mapping

pooling function

attention map

output mapping

# pooled vectors

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: "Universal" (Network & Settings)

| METHOD | EP | RESNET-50 | CONVNEXT-S | VIT-S | VIT-B |
|---|---|---|---|---|---|
| Baseline | 100 | 77.4 | 81.1 | 72.7 | 74.1 |
| CaiT | 100 | 77.3 | 81.2 | 72.6 | - |
| Slot | 100 | 77.3 | 80.9 | 72.9 | - |
| GE | 100 | 77.6 | 81.3 | 72.6 | - |
| SimPool | 100 | **78.0** | **81.7** | **74.3** | **75.1** |
| Baseline | 300 | 78.1† | 83.1 | 77.9 | - |
| SimPool | 300 | **78.7**† | **83.5** | **78.7** | - |

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: "Universal" (Network & Settings)

| METHOD | EP | RESNET-50 | CONVNEXT-S | VIT-S | VIT-B |
|---|---|---|---|---|---|
| Baseline | 100 | 77.4 | 81.1 | 72.7 | 74.1 |
| CaiT | 100 | 77.3 | 81.2 | 72.6 | - |
| Slot | 100 | 77.3 | 80.9 | 72.9 | - |
| GE | 100 | 77.6 | 81.3 | 72.6 | - |
| SimPool | 100 | **78.0** | **81.7** | **74.3** | **75.1** |
| Baseline | 300 | 78.1$^\dagger$ | 83.1 | 77.9 | - |
| SimPool | 300 | **78.7**$^\dagger$ | **83.5** | **78.7** | - |

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: "Universal" (Network & Settings)

| Method | Ep | ResNet-50 | ConvNext-S | ViT-S | ViT-B |
|---|---|---|---|---|---|
| Baseline | 100 | 77.4 | 81.1 | 72.7 | 74.1 |
| CaiT | 100 | 77.3 | 81.2 | 72.6 | - |
| Slot | 100 | 77.3 | 80.9 | 72.9 | - |
| GE | 100 | 77.6 | 81.3 | 72.6 | - |
| SimPool | 100 | **78.0** | **81.7** | **74.3** | **75.1** |
| Baseline | 300 | 78.1$^{\dagger}$ | 83.1 | 77.9 | - |
| SimPool | 300 | **78.7**$^{\dagger}$ | **83.5** | **78.7** | - |

Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

| Method | Ep | ResNet-50 | | ConvNeXt-S | | ViT-S | |
|---|---|---|---|---|---|---|---|
| | | $k$-NN | Prob | $k$-NN | Prob | $k$-NN | Prob |
| Baseline | 100 | 61.8 | 63.0 | 65.1 | 68.2 | 68.9 | 71.5 |
| SimPool | 100 | **63.8** | **64.4** | **68.8** | **72.2** | **69.8** | **72.8** |

Classification accuracy on ImageNet-1k;
Self-supervised pre-training w/ DINO;
Baseline: GAP for convolutional, [CLS] for transformers.

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: “Universal” (Network & Settings)

| METHOD | EP | RESNET-50 | CONVNEXT-S | VIT-S | VIT-B |
|---|---|---|---|---|---|
| Baseline | 100 | 77.4 | 81.1 | 72.7 | 74.1 |
| CaiT | 100 | 77.3 | 81.2 | 72.6 | - |
| Slot | 100 | 77.3 | 80.9 | 72.9 | - |
| GE | 100 | 77.6 | 81.3 | 72.6 | - |
| SimPool | 100 | **78.0** | **81.7** | **74.3** | **75.1** |
| Baseline | 300 | 78.1† | 83.1 | 77.9 | - |
| SimPool | 300 | **78.7**† | **83.5** | **78.7** | - |

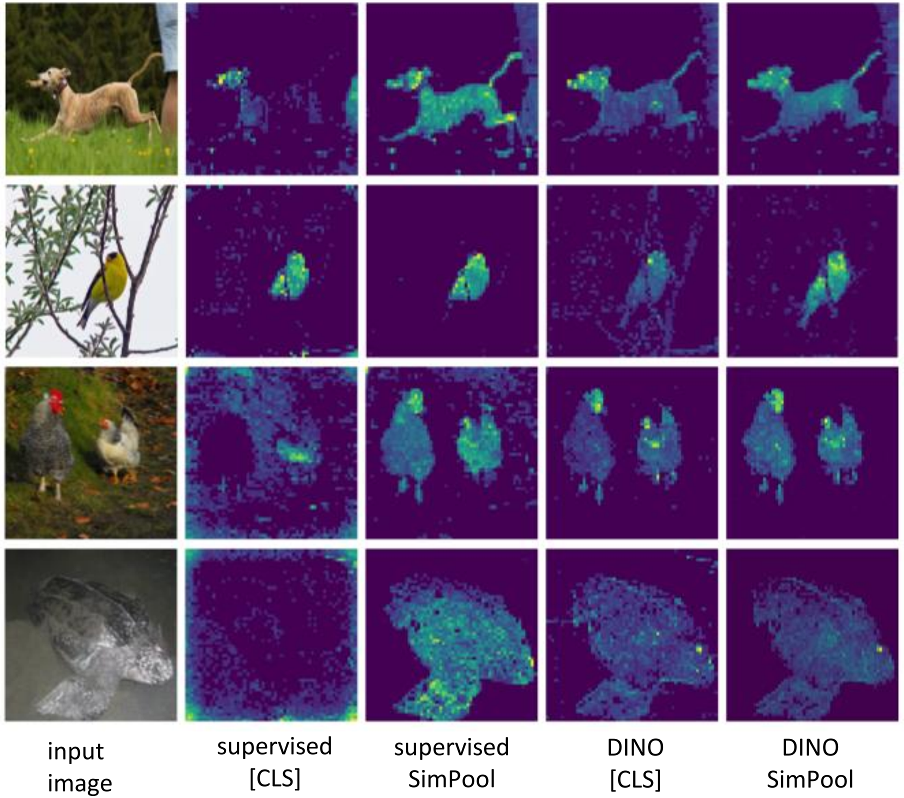
Classification accuracy on ImageNet-1k;
Supervised training;
Baseline: GAP for convolutional, [CLS] for transformers.

| METHOD | EP | RESNET-50 | | CONVNEXT-S | | VIT-S | |
|---|---|---|---|---|---|---|---|
| | | $k$-NN | PROB | $k$-NN | PROB | $k$-NN | PROB |
| Baseline | 100 | 61.8 | 63.0 | 65.1 | 68.2 | 68.9 | 71.5 |
| SimPool | 100 | **63.8** | **64.4** | **68.8** | **72.2** | **69.8** | **72.8** |

Classification accuracy on ImageNet-1k;
Self-supervised pre-training w/ DINO;
Baseline: GAP for convolutional, [CLS] for transformers.

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

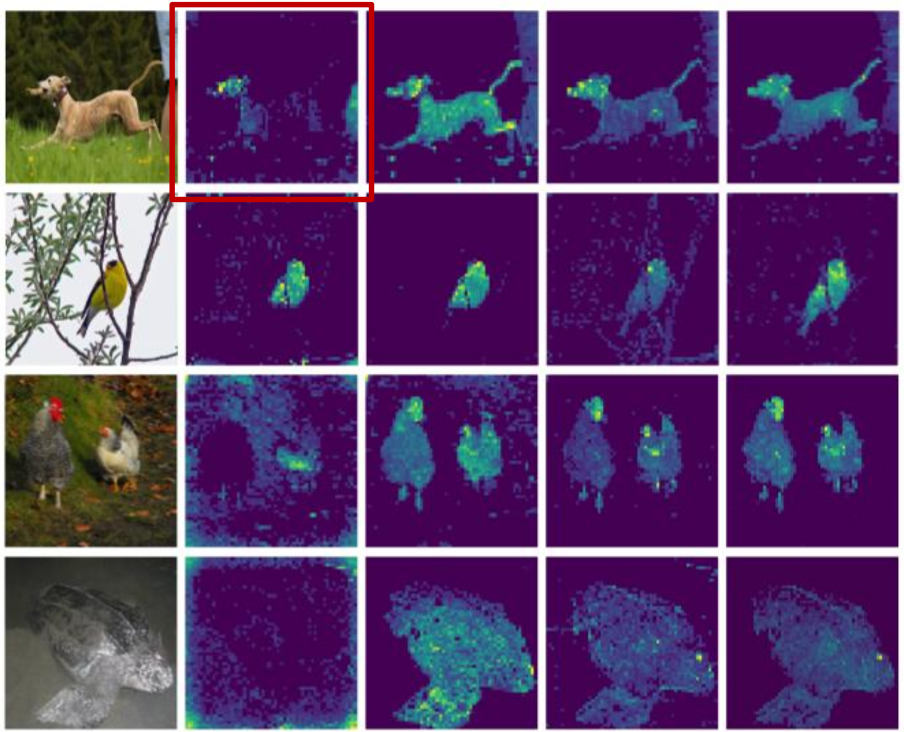
# Property: High-quality attention maps from Transformers




ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

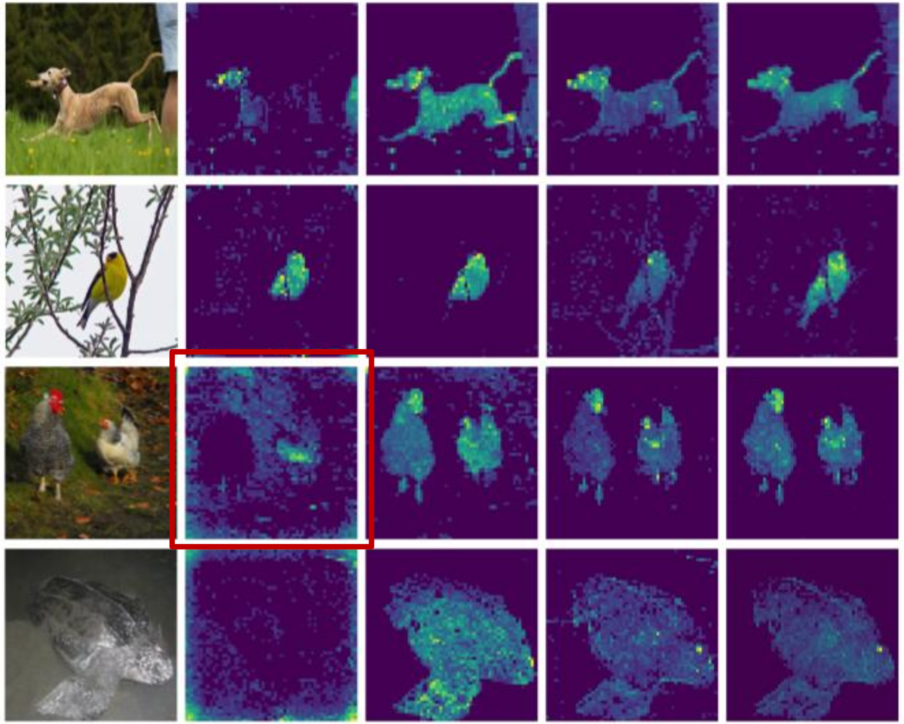
# Property: High-quality attention maps from Transformers




ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: High-quality attention maps from Transformers

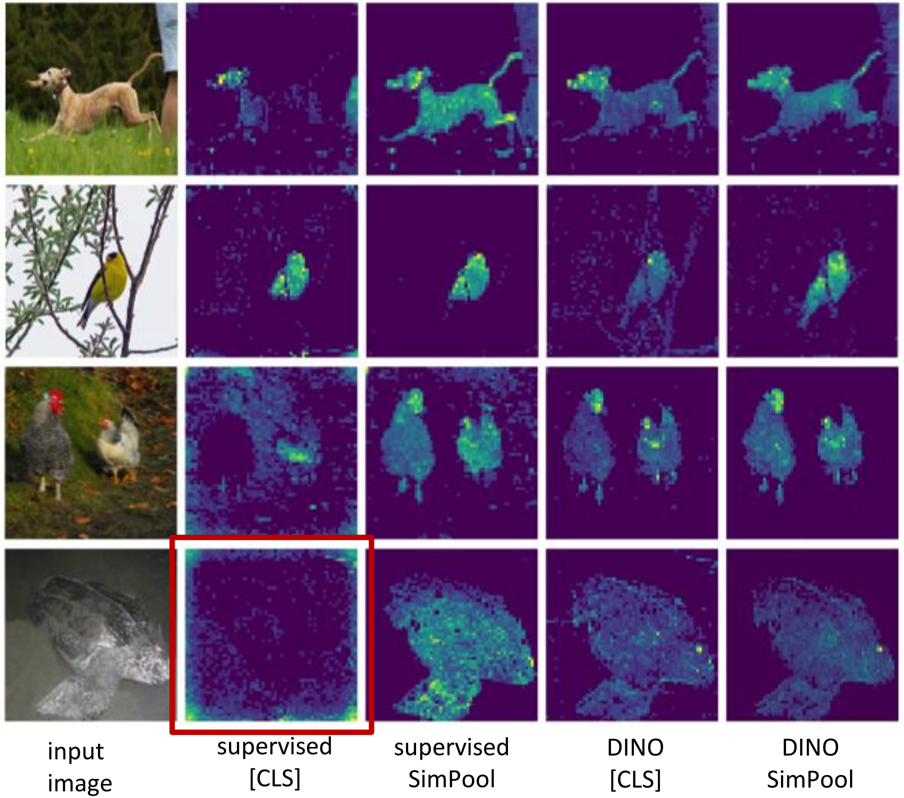


input image | supervised [CLS] | supervised SimPool | DINO [CLS] | DINO SimPool

ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

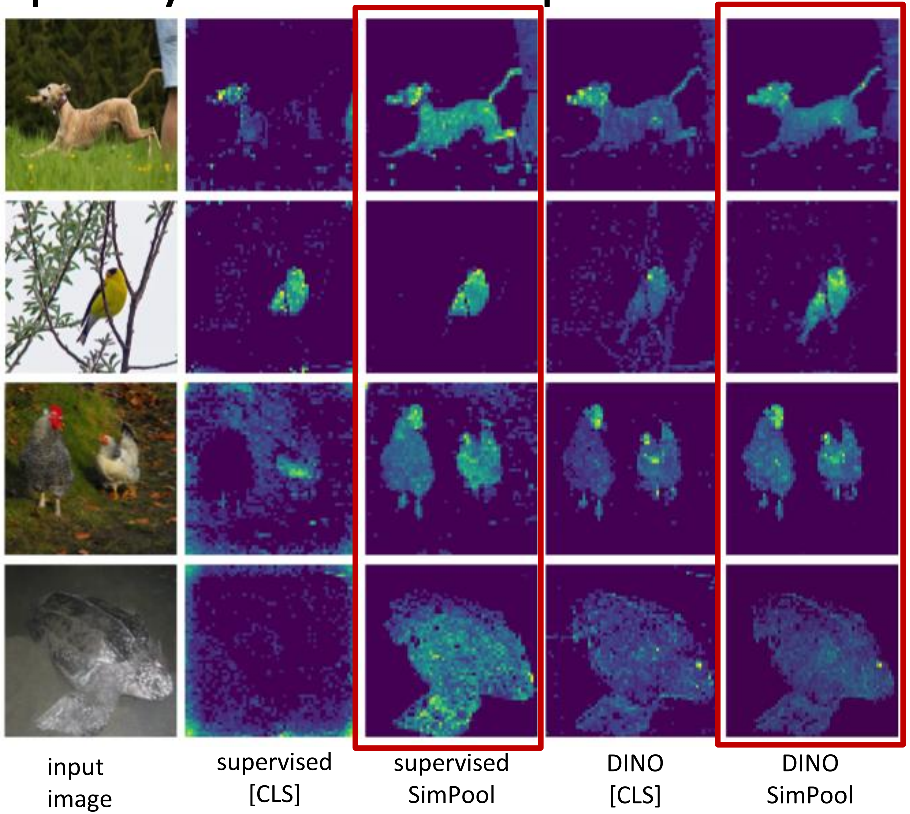
# Property: High-quality attention maps from Transformers




ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

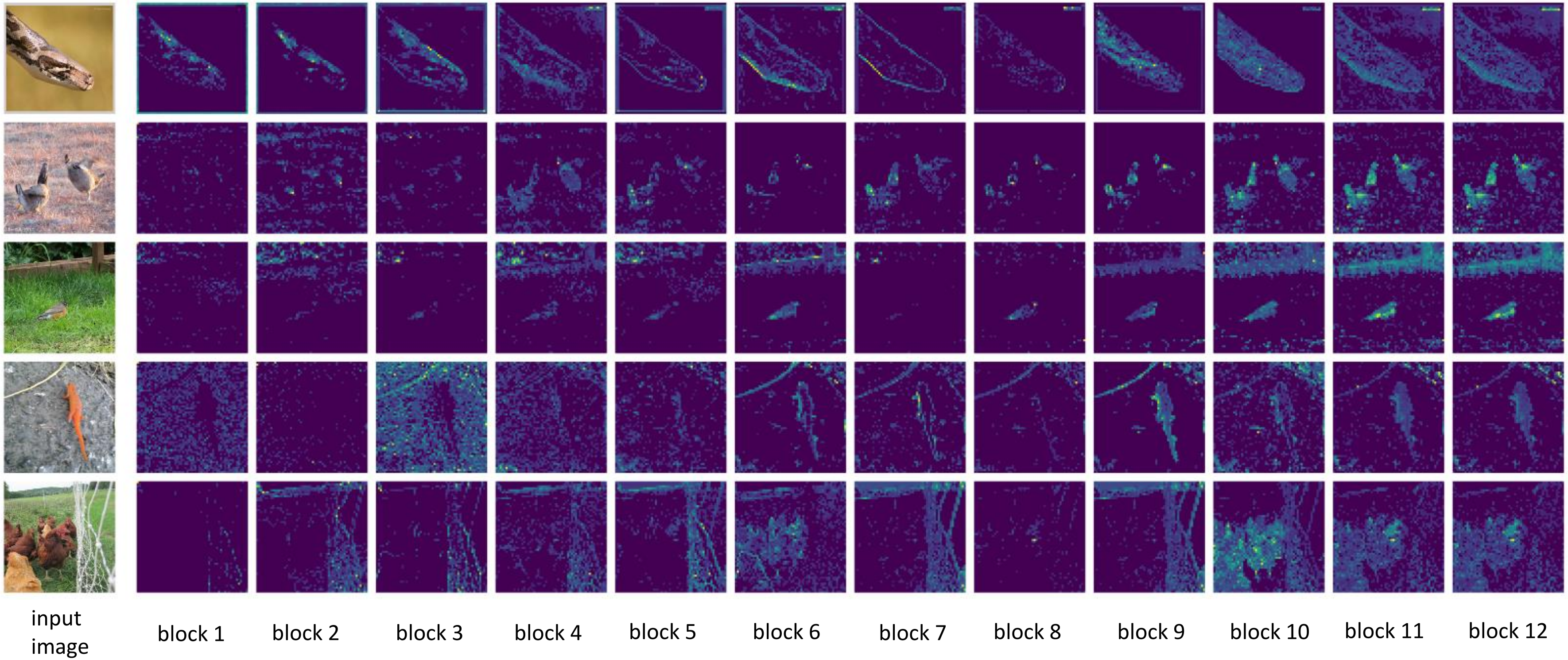
# Property: High-quality attention maps from Transformers




ViT-S on Imagenet-1k; mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

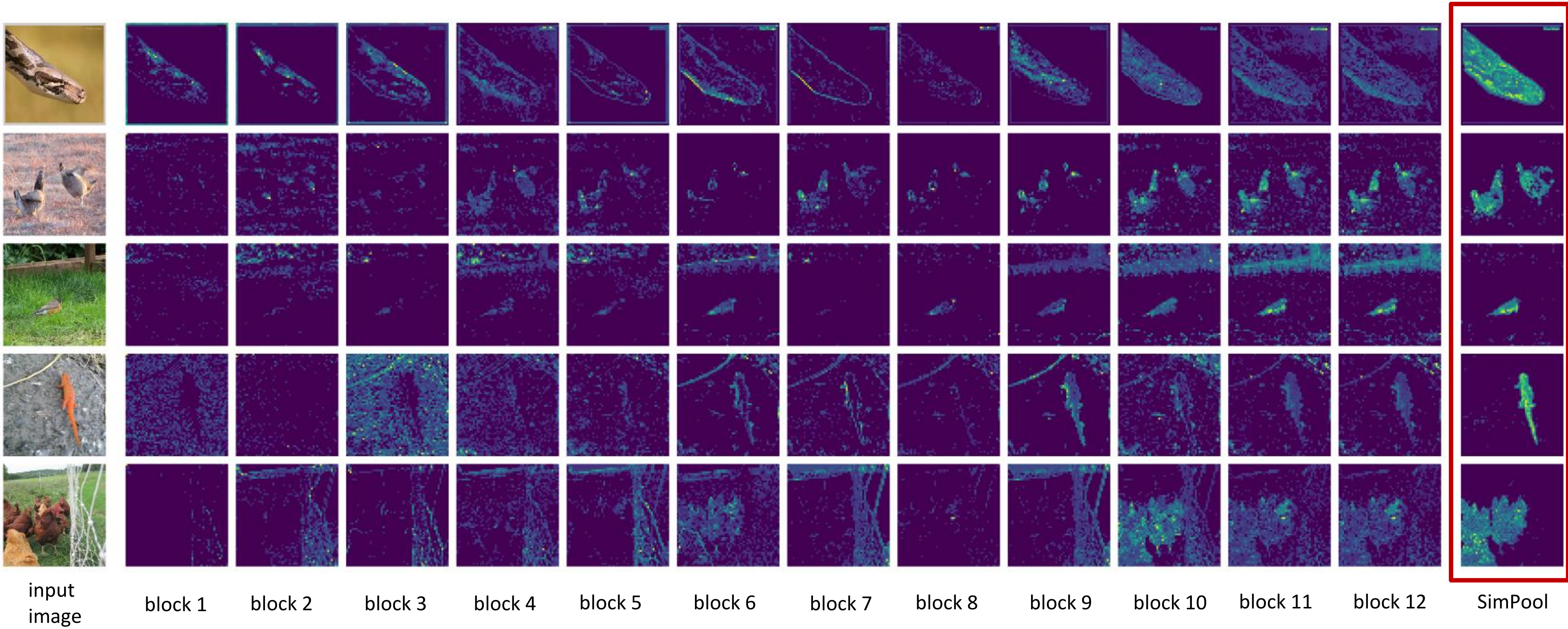
# Property: Resolving the attention "deficit"




ViT-S on Imagenet-1k; supervised training;
mean attention map of the [CLS]

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

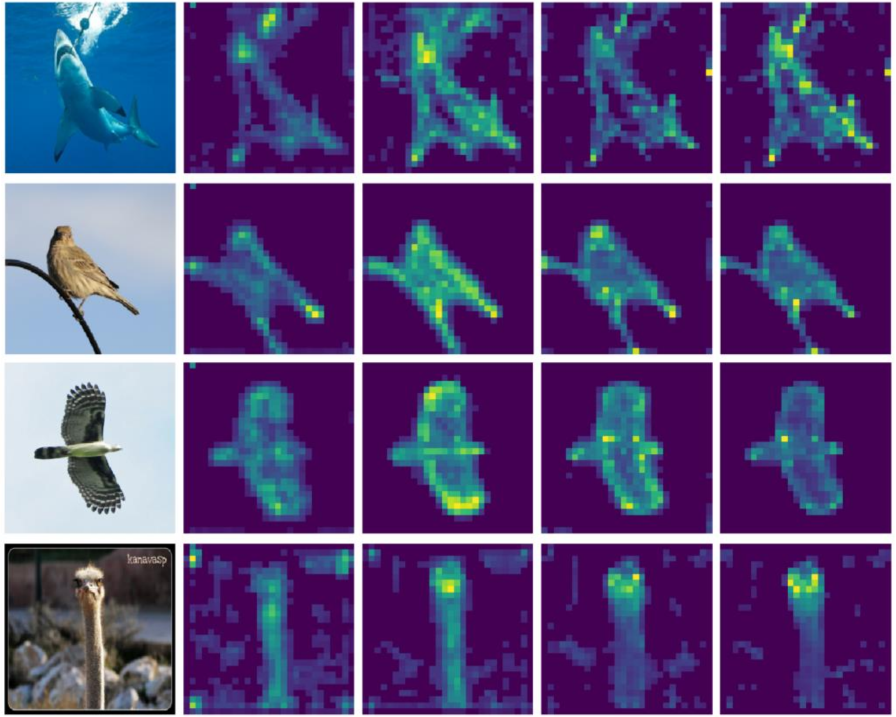
# Property: Resolving the attention "deficit"




ViT-S on Imagenet-1k; supervised training;
mean attention map of the [CLS] vs. SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: High-quality attention maps from CNNs




input image | ResNet-50 supervised | ResNet-50 DINO | ConvNeXt-S supervised | ConvNeXt-S DINO

ResNet-50, ConvNeXt-S on Imagenet-1k; supervised training; SimPool attention map

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: Localization

| Method | Supervised | | Self-Supervised | |
|---|---|---|---|---|
| | CUB | ImageNet | CUB | ImageNet |
| Baseline | 63.1 | 53.6 | 82.7 | 62.0 |
| SimPool | **77.9** | **64.4** | **86.1** | **66.1** |
| Baseline@20 | 62.4 | 50.5 | 65.5 | 52.5 |
| SimPool@20 | **74.0** | **62.6** | **72.5** | **58.7** |

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

# Property: Localization

| Method | Supervised | | Self-Supervised | |
|---|---|---|---|---|
| | CUB | ImageNet | CUB | ImageNet |
| Baseline | 63.1 | 53.6 | 82.7 | 62.0 |
| SimPool | **77.9** | **64.4** | **86.1** | **66.1** |
| Baseline@20 | 62.4 | 50.5 | 65.5 | 52.5 |
| SimPool@20 | **74.0** | **62.6** | **72.5** | **58.7** |

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

| Method | DINO-seg | | |
|---|---|---|---|
| | VOC07 | VOC12 | COCO |
| Baseline | 30.8 | 31.0 | 36.7 |
| SimPool | **53.2** | **56.2** | **43.4** |
| Baseline@20 | 14.9 | 14.8 | 19.9 |
| SimPool@20 | **49.2** | **54.8** | **37.9** |

Unsupervised object discovery CorLoc with ViT-S;
DINO-SEG uses attention maps;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

# Property: Localization

| Method | Supervised | | Self-Supervised | |
|---|---|---|---|---|
| | CUB | ImageNet | CUB | ImageNet |
| Baseline | 63.1 | 53.6 | 82.7 | 62.0 |
| SimPool | **77.9** | **64.4** | **86.1** | **66.1** |
| Baseline@20 | 62.4 | 50.5 | 65.5 | 52.5 |
| SimPool@20 | **74.0** | **62.6** | **72.5** | **58.7** |

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

| Method | DINO-seg | | | LOST | | |
|---|---|---|---|---|---|---|
| | VOC07 | VOC12 | COCO | VOC07 | VOC12 | COCO |
| Baseline | 30.8 | 31.0 | 36.7 | 55.5 | 59.4 | 46.6 |
| SimPool | **53.2** | **56.2** | **43.4** | **59.8** | **65.0** | **49.4** |
| Baseline@20 | 14.9 | 14.8 | 19.9 | 50.7 | 56.6 | 40.9 |
| SimPool@20 | **49.2** | **54.8** | **37.9** | **53.9** | **58.8** | **46.1** |

Unsupervised object discovery CorLoc with ViT-S;
DINO-seg uses attention maps;
LOST uses raw features;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020
Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

# Property: Localization

| Method | Supervised | | Self-Supervised | |
|---|---|---|---|---|
| | CUB | ImageNet | CUB | ImageNet |
| Baseline | 63.1 | 53.6 | 82.7 | 62.0 |
| SimPool | **77.9** | **64.4** | **86.1** | **66.1** |
| Baseline@20 | 62.4 | 50.5 | 65.5 | 52.5 |
| SimPool@20 | **74.0** | **62.6** | **72.5** | **58.7** |

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

- ✓ Up to +14% when supervised and up to +7% when self-supervised

| Method | DINO-seg | | | LOST | | |
|---|---|---|---|---|---|---|
| | VOC07 | VOC12 | COCO | VOC07 | VOC12 | COCO |
| Baseline | 30.8 | 31.0 | 36.7 | 55.5 | 59.4 | 46.6 |
| SimPool | **53.2** | **56.2** | **43.4** | **59.8** | **65.0** | **49.4** |
| Baseline@20 | 14.9 | 14.8 | 19.9 | 50.7 | 56.6 | 40.9 |
| SimPool@20 | **49.2** | **54.8** | **37.9** | **53.9** | **58.8** | **46.1** |

Unsupervised object discovery CorLoc with ViT-S;
DINO-SEG uses attention maps;
LOST uses raw features;
@20: at epoch 20

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020
Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

# Property: Localization

| Method | Supervised | | Self-Supervised | |
|---|---|---|---|---|
| | CUB | ImageNet | CUB | ImageNet |
| Baseline | 63.1 | 53.6 | 82.7 | 62.0 |
| SimPool | **77.9** | **64.4** | **86.1** | **66.1** |
| Baseline@20 | 62.4 | 50.5 | 65.5 | 52.5 |
| SimPool@20 | **74.0** | **62.6** | **72.5** | **58.7** |

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20

✓ Up to +14% when supervised and up to +7% when self-supervised

| Method | DINO-seg | | | LOST | | |
|---|---|---|---|---|---|---|
| | VOC07 | VOC12 | COCO | VOC07 | VOC12 | COCO |
| Baseline | 30.8 | 31.0 | 36.7 | 55.5 | 59.4 | 46.6 |
| SimPool | **53.2** | **56.2** | **43.4** | **59.8** | **65.0** | **49.4** |
| Baseline@20 | 14.9 | 14.8 | 19.9 | 50.7 | 56.6 | 40.9 |
| SimPool@20 | **49.2** | **54.8** | **37.9** | **53.9** | **58.8** | **46.1** |

Unsupervised object discovery CorLoc with ViT-S;
DINO-SEG uses attention maps;
LOST uses raw features;
@20: at epoch 20

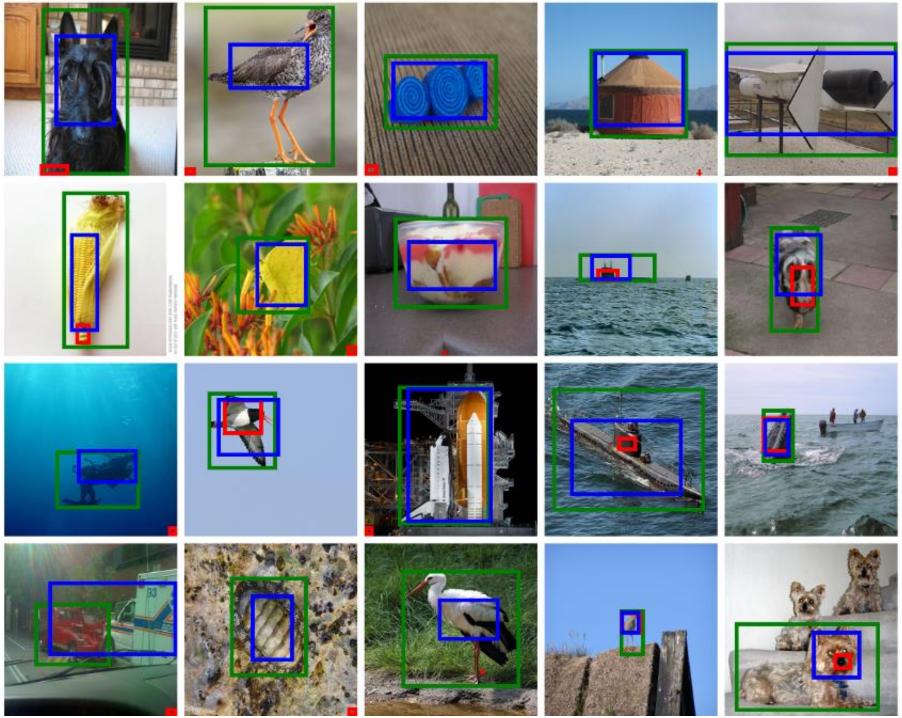
✓ Up to +25% for DINO-seg and up to +6% for LOST

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020
Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021
Simeoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC 2021

# Property: Localization

| Method | Supervised | | Self-Supervised | |
|---|---|---|---|---|
| | CUB | ImageNet | CUB | ImageNet |
| Baseline | 63.1 | 53.6 | 82.7 | 62.0 |
| SimPool | **77.9** | **64.4** | **86.1** | **66.1** |
| Baseline@20 | 62.4 | 50.5 | 65.5 | 52.5 |
| SimPool@20 | **74.0** | **62.6** | **72.5** | **58.7** |

Object localization MaxBoxAccV2 with ViT-S;
Baseline: mean attention map of the [CLS];
SimPool attention map;
@20: at epoch 20



Object localization on ImageNet-1k;
green: ground-truth; red: baseline; blue: SimPool

Choe et al., Evaluating weakly supervised object localization methods right, CVPR 2020

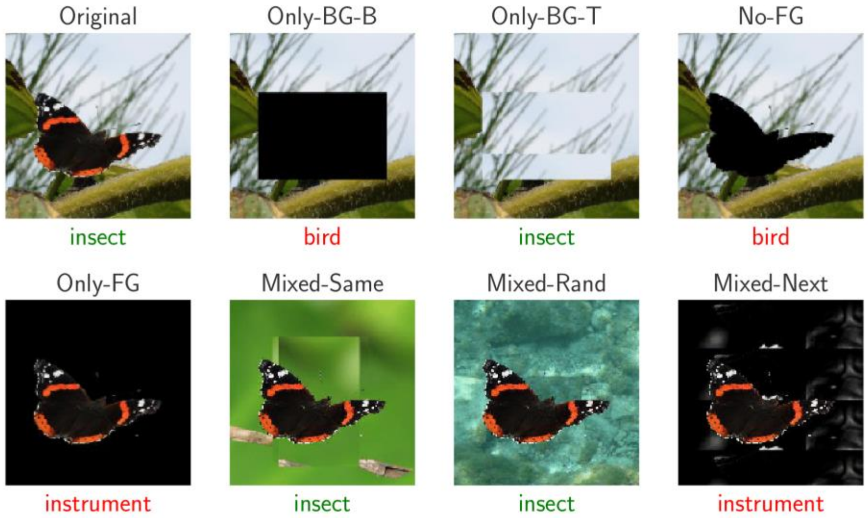
Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Property: Background robustness

| METHOD | OF | MS | MR | MN | NF | OBB | OBT | IN-9 |
|---|---|---|---|---|---|---|---|---|
| SUPERVISED | | | | | | | | |
| Baseline | 66.4 | 79.1 | 67.4 | 65.5 | 37.2 | 12.9 | 15.2 | 92.0 |
| SimPool | **71.8** | **80.2** | **69.3** | **67.3** | **42.8** | **15.2** | **15.6** | **92.9** |
| SELF-SUPERVISED + LINEAR PROBING | | | | | | | | |
| Baseline | **87.3** | 87.9 | 78.5 | 76.7 | 47.9 | **20.0** | **16.9** | 95.3 |
| SimPool | **87.3** | **88.1** | **80.6** | **78.7** | **48.2** | 17.8 | 16.7 | **95.6** |

Background robustness

Classification accuracy on IN-9 with ViT-S




Classification robustness against background changes

Xiao et al., Noise or Signal: The Role of Image backgrounds in Object Recognition; ICLR 2021

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

| NETWORK | POOLING | DEPTH | INIT | ACCURACY | #PARAMS |
|---|---|---|---|---|---|
| BASE | GAP | 12 | 12 | 73.3 | 22.1M |
| BASE | CLS | 12 | 0 | 72.7 | 22.1M |
| BASE + 1 | | 13 | 0 | 73.2 | 23.8M |
| BASE + 2 | | 14 | 0 | 73.7 | 25.6M |
| BASE + 3 | | 15 | 0 | 73.8 | 27.4M |
| BASE + 4 | | 16 | 0 | 73.9 | 29.2M |
| BASE + 5 | | 17 | 0 | **74.6** | 30.9M |
| BASE | SimPool | 12 | 12 | **74.3** | 22.3M |
| BASE − 1 | | 11 | 11 | 73.9 | 20.6M |
| BASE − 2 | | 10 | 10 | 73.6 | 18.7M |
| BASE − 3 | | 9 | 9 | 72.5 | 17.0M |

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

Add ViT blocks when using [CLS]

| NETWORK | POOLING | DEPTH | INIT | ACCURACY | #PARAMS |
|---|---|---|---|---|---|
| BASE | GAP | 12 | 12 | 73.3 | 22.1M |
| BASE | CLS | 12 | 0 | 72.7 | 22.1M |
| BASE + 1 | CLS | 13 | 0 | 73.2 | 23.8M |
| BASE + 2 | CLS | 14 | 0 | 73.7 | 25.6M |
| BASE + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| BASE + 4 | CLS | 16 | 0 | 73.9 | 29.2M |
| BASE + 5 | CLS | 17 | 0 | **74.6** | 30.9M |
| BASE | SimPool | 12 | 12 | **74.3** | 22.3M |
| BASE − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| BASE − 2 | SimPool | 10 | 10 | 73.6 | 18.7M |
| BASE − 3 | SimPool | 9 | 9 | 72.5 | 17.0M |

Classification accuracy of ViT-S on ImageNet-1k; Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

Add ViT blocks when using [CLS]

| Network | Pooling | Depth | Init | Accuracy | #params |
|---|---|---|---|---|---|
| Base | GAP | 12 | 12 | 73.3 | 22.1M |
| Base | CLS | 12 | 0 | 72.7 | 22.1M |
| Base + 1 | CLS | 13 | 0 | 73.2 | 23.8M |
| Base + 2 | CLS | 14 | 0 | 73.7 | 25.6M |
| Base + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| Base + 4 | CLS | 16 | 0 | 73.9 | 29.2M |
| Base + 5 | CLS | 17 | 0 | **74.6** | 30.9M |
| Base | SimPool | 12 | 12 | **74.3** | 22.3M |
| Base − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| Base − 2 | SimPool | 10 | 10 | 73.6 | 18.7M |
| Base − 3 | SimPool | 9 | 9 | 72.5 | 17.0M |

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

Add ViT blocks when using [CLS]

| NETWORK | POOLING | DEPTH | INIT | ACCURACY | #PARAMS |
|---|---|---|---|---|---|
| BASE | GAP | 12 | 12 | 73.3 | 22.1M |
| BASE | CLS | 12 | 0 | 72.7 | 22.1M |
| BASE + 1 | | 13 | 0 | 73.2 | 23.8M |
| BASE + 2 | | 14 | 0 | 73.7 | 25.6M |
| BASE + 3 | | 15 | 0 | 73.8 | 27.4M |
| BASE + 4 | | 16 | 0 | 73.9 | 29.2M |
| BASE + 5 | | 17 | 0 | **74.6** | 30.9M |
| BASE | SimPool | 12 | 12 | **74.3** | 22.3M |
| BASE − 1 | | 11 | 11 | 73.9 | 20.6M |
| BASE − 2 | | 10 | 10 | 73.6 | 18.7M |
| BASE − 3 | | 9 | 9 | 72.5 | 17.0M |

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

Add ViT blocks when using [CLS]

| Network | Pooling | Depth | Init | Accuracy | #Params |
|---|---|---|---|---|---|
| Base | GAP | 12 | 12 | 73.3 | 22.1M |
| Base | CLS | 12 | 0 | 72.7 | 22.1M |
| Base + 1 | CLS | 13 | 0 | 73.2 | 23.8M |
| Base + 2 | CLS | 14 | 0 | 73.7 | 25.6M |
| Base + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| Base + 4 | CLS | 16 | 0 | 73.9 | 29.2M |
| Base + 5 | CLS | 17 | 0 | **74.6** | 30.9M |
| Base | SimPool | 12 | 12 | **74.3** | 22.3M |
| Base − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| Base − 2 | SimPool | 10 | 10 | 73.6 | 18.7M |
| Base − 3 | SimPool | 9 | 9 | 72.5 | 17.0M |

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

Add ViT blocks when using [CLS]

| Network | Pooling | Depth | Init | Accuracy | #Params |
|---|---|---|---|---|---|
| Base | GAP | 12 | 12 | 73.3 | 22.1M |
| Base |  | 12 | 0 | 72.7 | 22.1M |
| Base + 1 |  | 13 | 0 | 73.2 | 23.8M |
| Base + 2 |  | 14 | 0 | 73.7 | 25.6M |
| Base + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| Base + 4 |  | 16 | 0 | 73.9 | 29.2M |
| Base + 5 |  | 17 | 0 | **74.6** | 30.9M |
| Base |  | 12 | 12 | **74.3** | 22.3M |
| Base − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| Base − 2 |  | 10 | 10 | 73.6 | 18.7M |
| Base − 3 |  | 9 | 9 | 72.5 | 17.0M |

5 extra blocks or >8M more parameters to exceed!

Classification accuracy of ViT-S on ImageNet-1k; Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

| Network | Pooling | Depth | Init | Accuracy | #params |
|---|---|---|---|---|---|
| Base | GAP | 12 | 12 | 73.3 | 22.1M |
| Base | CLS | 12 | 0 | 72.7 | 22.1M |
| Base + 1 | | 13 | 0 | 73.2 | 23.8M |
| Base + 2 | | 14 | 0 | 73.7 | 25.6M |
| Base + 3 | | 15 | 0 | 73.8 | 27.4M |
| Base + 4 | | 16 | 0 | 73.9 | 29.2M |
| Base + 5 | | 17 | 0 | **74.6** | 30.9M |
| Base | SimPool | 12 | 12 | **74.3** | 22.3M |
| Base − 1 | | 11 | 11 | 73.9 | 20.6M |
| Base − 2 | | 10 | 10 | 73.6 | 18.7M |
| Base − 3 | | 9 | 9 | 72.5 | 17.0M |

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

| Network | Pooling | Depth | Init | Accuracy | #Params |
|---|---|---|---|---|---|
| Base | GAP | 12 | 12 | 73.3 | 22.1M |
| Base | CLS | 12 | 0 | 72.7 | 22.1M |
| Base + 1 | CLS | 13 | 0 | 73.2 | 23.8M |
| Base + 2 | CLS | 14 | 0 | 73.7 | 25.6M |
| Base + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| Base + 4 | CLS | 16 | 0 | 73.9 | 29.2M |
| Base + 5 | CLS | 17 | 0 | **74.6** | 30.9M |
| Base | SimPool | 12 | 12 | **74.3** | 22.3M |
| Base − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| Base − 2 | SimPool | 10 | 10 | 73.6 | 18.7M |
| Base − 3 | SimPool | 9 | 9 | 72.5 | 17.0M |

Remove ViT blocks when using SimPool

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

| NETWORK | POOLING | DEPTH | INIT | ACCURACY | #PARAMS |
|---|---|---|---|---|---|
| BASE | GAP | 12 | 12 | 73.3 | 22.1M |
| BASE | CLS | 12 | 0 | 72.7 | 22.1M |
| BASE + 1 | CLS | 13 | 0 | 73.2 | 23.8M |
| BASE + 2 | CLS | 14 | 0 | 73.7 | 25.6M |
| BASE + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| BASE + 4 | CLS | 16 | 0 | 73.9 | 29.2M |
| BASE + 5 | CLS | 17 | 0 | **74.6** | 30.9M |
| BASE | SimPool | 12 | 12 | **74.3** | 22.3M |
| BASE − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| BASE − 2 | SimPool | 10 | 10 | 73.6 | 18.7M |
| BASE − 3 | SimPool | 9 | 9 | 72.5 | 17.0M |

Remove ViT blocks when using SimPool

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Performance vs. Parameters

| Network | Pooling | Depth | Init | Accuracy | #params |
|---|---|---|---|---|---|
| Base | GAP | 12 | 12 | 73.3 | 22.1M |
| Base | | 12 | 0 | 72.7 | 22.1M |
| Base + 1 | | 13 | 0 | 73.2 | 23.8M |
| Base + 2 | | 14 | 0 | 73.7 | 25.6M |
| Base + 3 | CLS | 15 | 0 | 73.8 | 27.4M |
| Base + 4 | | 16 | 0 | 73.9 | 29.2M |
| Base + 5 | | 17 | 0 | **74.6** | 30.9M |
| Base | | 12 | 12 | **74.3** | 22.3M |
| Base − 1 | SimPool | 11 | 11 | 73.9 | 20.6M |
| Base − 2 | | 10 | 10 | 73.6 | 18.7M |
| Base − 3 | | 9 | 9 | 72.5 | 17.0M |

Remove ViT blocks when using SimPool

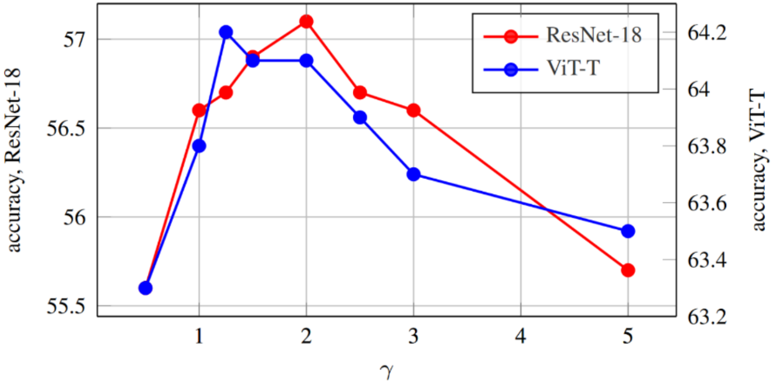
3 less blocks or 5M less parameters to be on par!

Classification accuracy of ViT-S on ImageNet-1k;
Supervised training;

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# The effect of γ

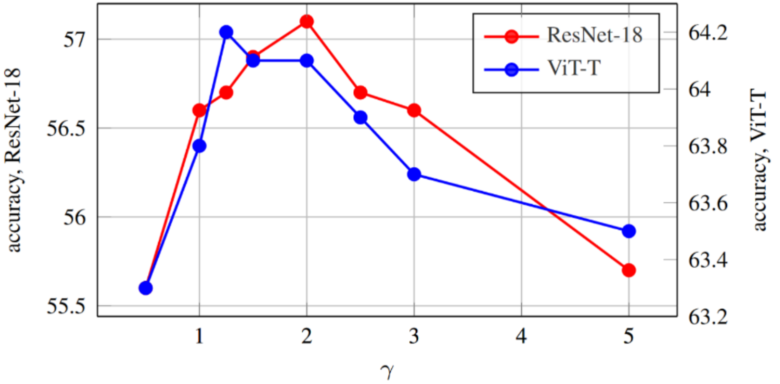
γ is a hyperparameter!

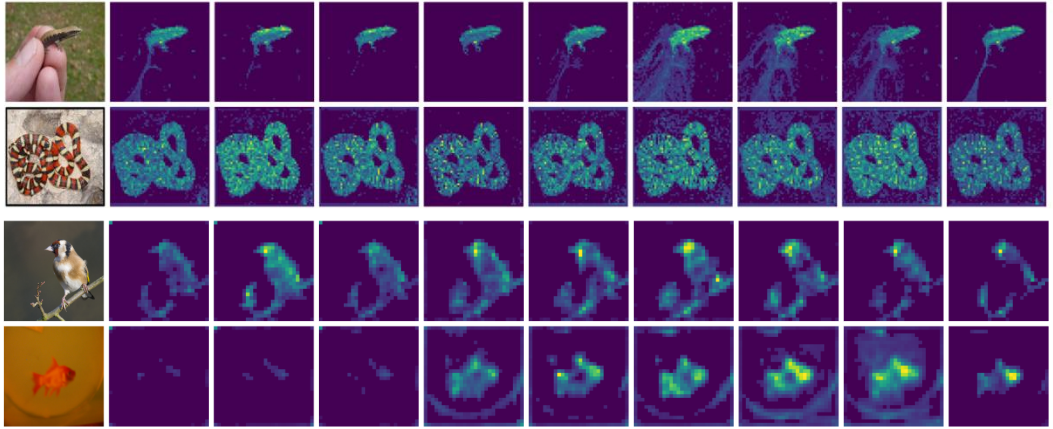


Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# The effect of γ

γ is a hyperparameter!



ViT

ResNet

input image | no γ | γ=0.5 | γ=1.0 | γ=1.25 | γ=1.5 | γ=2.0 | γ=2.5 | γ=3.0 | γ=5.0

# Conclusion

SimPool:

- ✓ Improves performance of convolutional networks and transformers under supervised or self-supervised setting
- ✓ Outperforms the other pooling methods
- ✓ Incurs low additional cost
- ✓ Produces high-quality attention maps that delineate object boundaries
- ✓ Presents strong localization properties

Psomas et al., Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?, ICCV 2023

# Collaborators

Ioannis
Kakogeorgiou

Spyros
Gidaris

Andrei
Bursuc

Konstantinos
Karantzalos

Yannis
Avrithis

Nikos
Komodakis