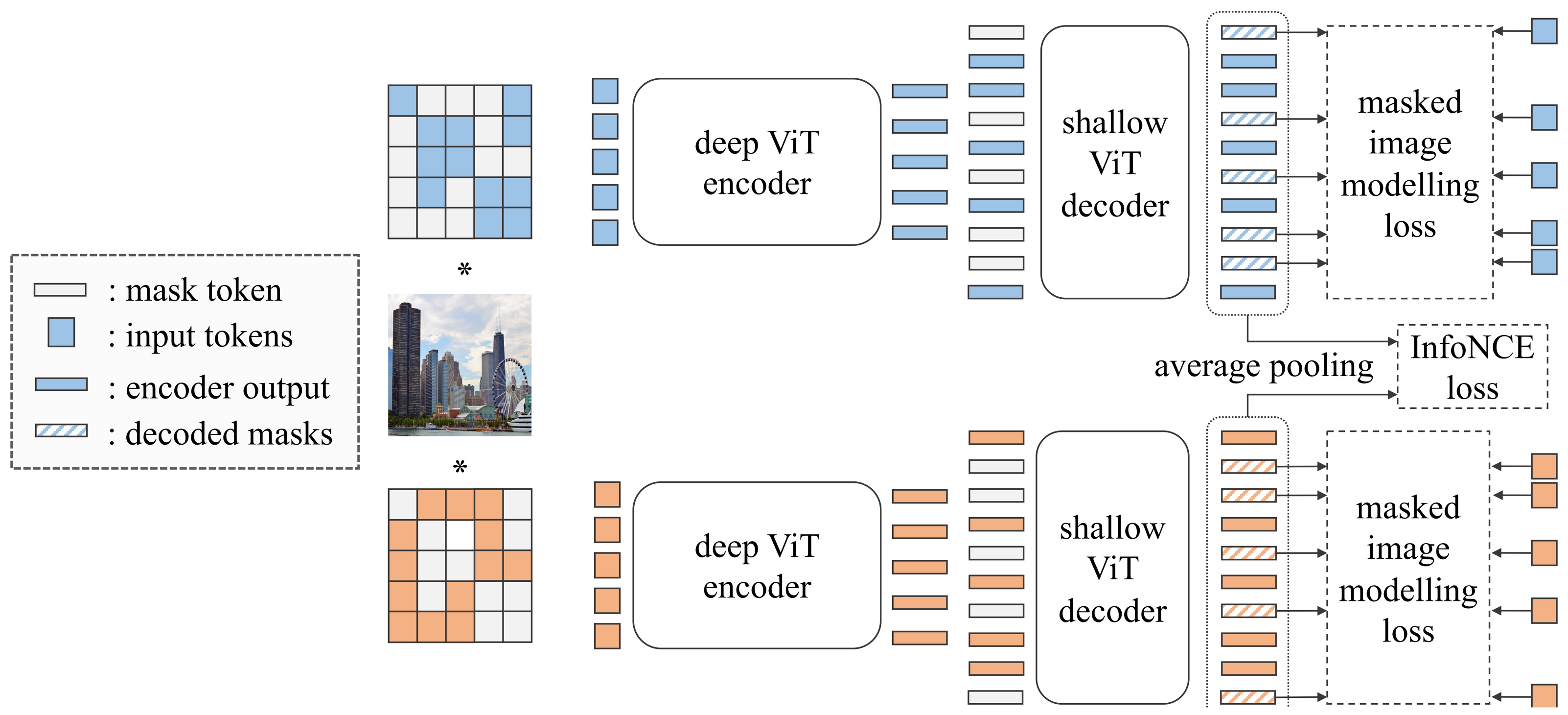
Is Large-scale Pre-training always Necessary for Vision Transformers?

Alaaeldin El-Nouby*,[1,2] Gautier Izacard*,[1,2] Hugo Touvron[1,3]
Ivan Laptev[2] Hervé Jégou[1] Edouard Grave[1]
[1]Meta AI
[2]Inria
[3]Sorbonne University

SplitMask



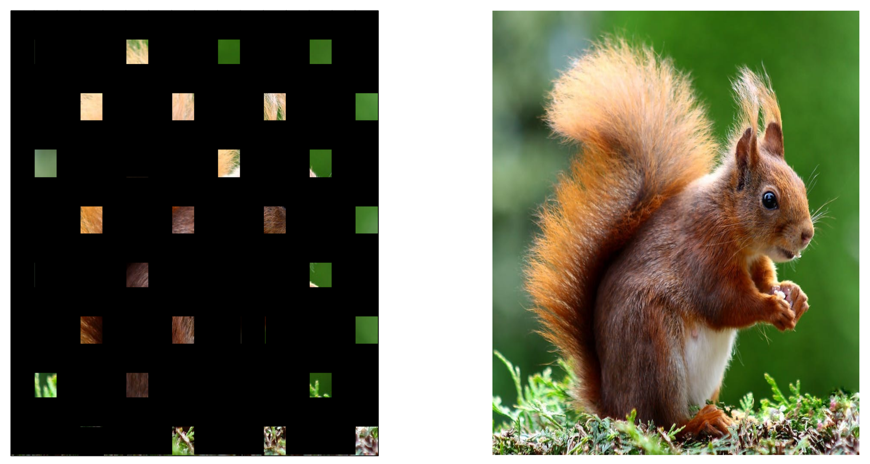
SplitMask is a denoising autoencoder that consists of three steps:

- *Split*: input image is split into 2 disjoint subsets and processed separately with a ViT encoder with shared parameters.
- *Inpaint*: the output is processed with a shallow decoder to predict missing patches in each branch.
- *Match*: The decoder outputs from both branches are contrastively trained to increase the similarity between their global descriptors (via Average pooling)

Method	Split	Inpaint	Match	Finetune	Lin.	Hours
BEiT [1]	✗	✓	✗	82.8	41.0	32.5
SplitMask	✓	✓	✗	83.3	46.4	**31.0**
	✓	✗	✓	79.3	4.0	32.5
	✓	✓	✓	**83.6**	**46.5**	34.0

Hypothesis

- Denoising Autoencoding methods (e.g. SplitMask, BEiT) are more sample efficient compared to joint embedding methods.



- Denoising Autoencoding methods are more robust to change in pre-training dataset nautre. They can be trained effectively using non-object centric datasets.

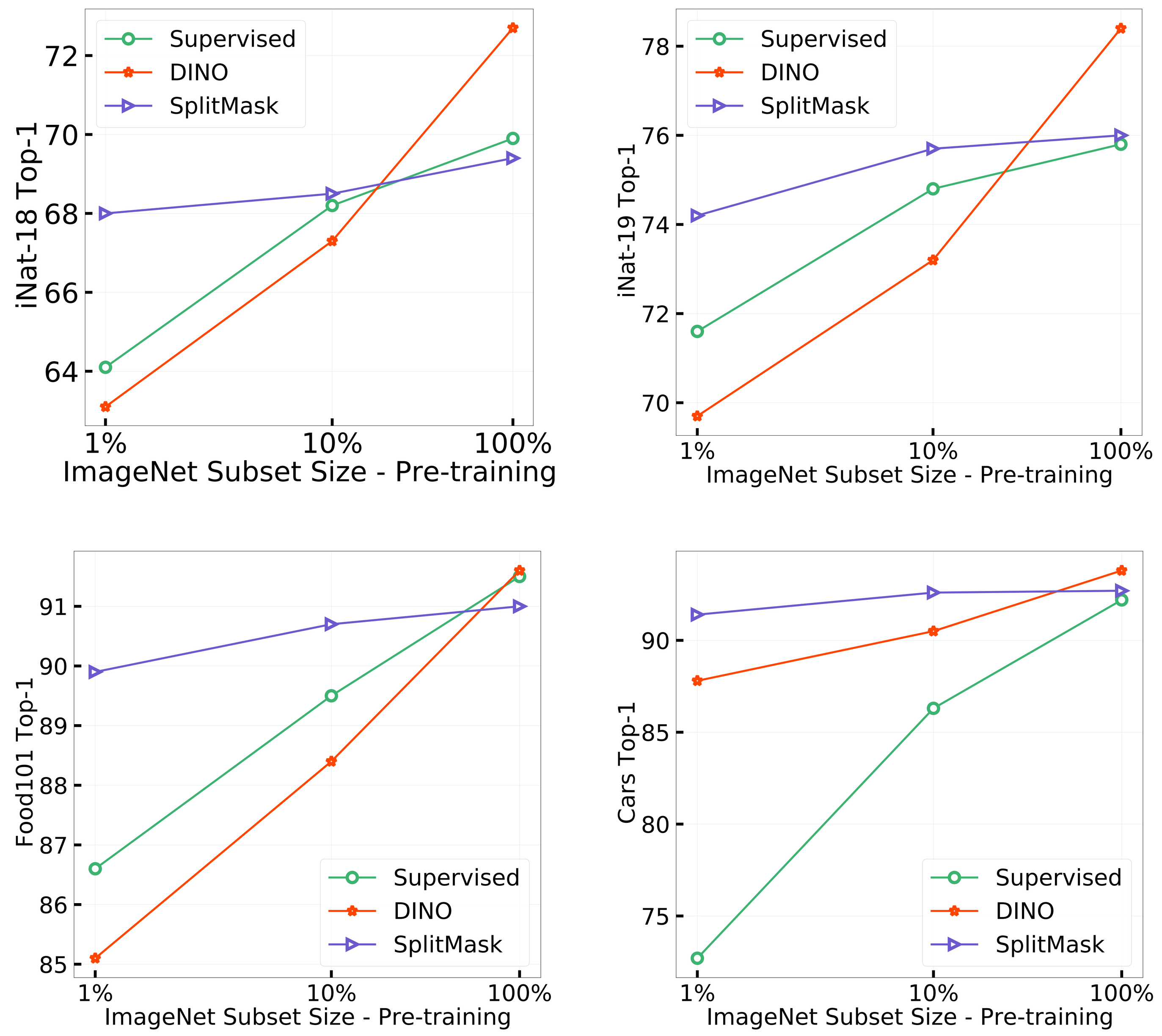
Method	IMNet 1% *epochs: 30k*	IMNet 10% *epochs: 3k*	IMNet Full *epochs: 300*	COCO *epochs: 3k*
Supervised	71.6	75.0	75.8	_
DINO [2]	70.1	73.1	**78.4**	71.9
BEiT [1]	74.1	74.5	75.2	74.4
SplitMask	**74.8**	**75.4**	75.4	**76.3**

Visual word targets with no training

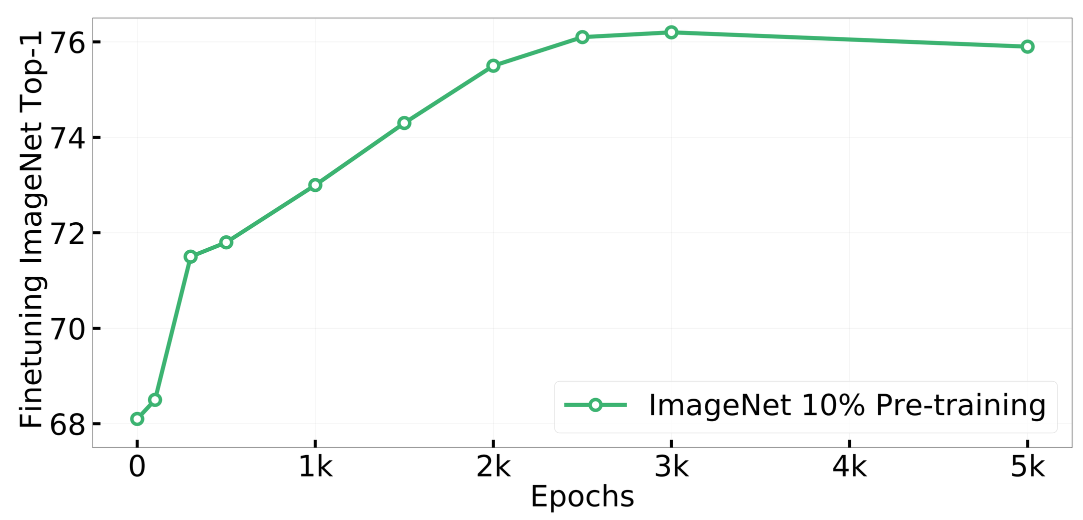
- Multiple simple methods that require no training can be used to generate the per-patch visual words, eliminating the need for pre-trained dVAE.

	DALL-E	Rand. Proj.	Rand. Patches	K-Means
iNat19	75.2	75.2	75.3	75.0

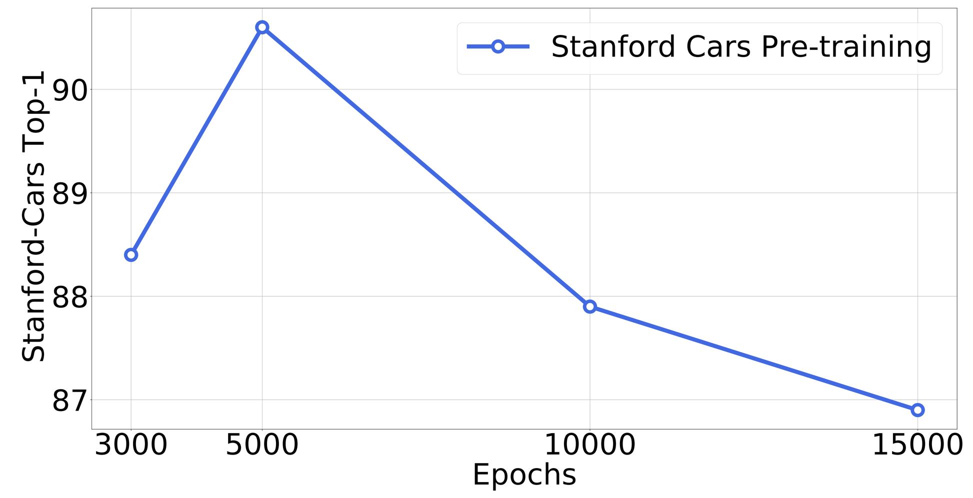
Sample Efficiency



How long to pre-train ?



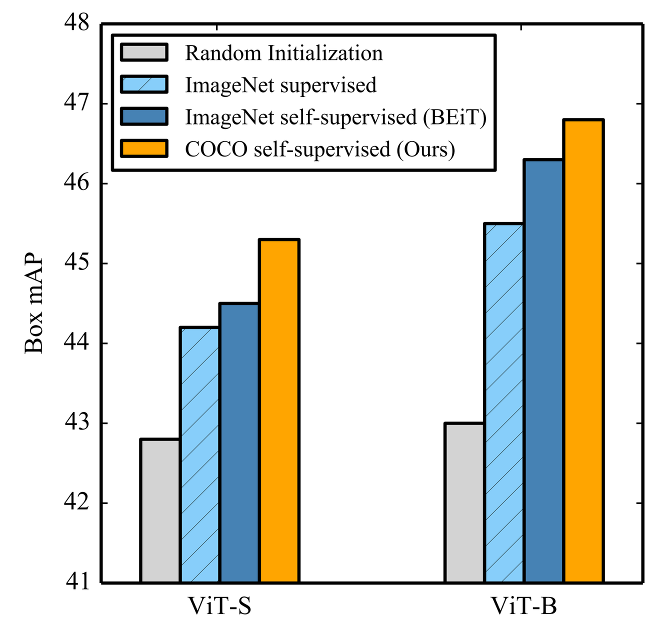
Is training for longer always better ?



Results

SplitMask provides performance on par, and in some cases improved, when pre-trained on the target dataset indicating that denoising autoencoding methods do not rely on large scale datasets for successful pre-training.

COCO Object Detection



ADE20k Semantic Segmentation

Method	Pre-training Supervised	Pre-training IMNet	Pre-training ADE20k	mIoU
Random Init.	✗	✗	✗	25.4
DeiT [3]	✓	✗	✗	46.1
BEiT [1]	✗	✓	✗	45.6
BEiT	✗	✗	✓	45.6
SplitMask	✗	✗	✓	45.7

Classification

Method	Backbone	Supervised pre-training	Data Used IMNet	Data Used Target	iNat-18 437k	iNat-19 265k	Food 101 75k	Cars 8k
Random Init.	ViT-S	✗	✗	✓	59.6	67.5	84.7	35.3
DeiT [3]		✓	✓	✓	<u>69.9</u>	75.8	**91.5**	92.2
BEiT [1]		✗	✓	✓	68.1	75.2	90.5	92.4
BEiT		✗	✗	✓	68.8	<u>76.1</u>	90.7	<u>92.7</u>
SplitMask		✗	✗	✓	**70.1**	**76.3**	**91.5**	**92.8**
Random Init.	ViT-B	✗	✗	✓	59.6	68.1	83.3	36.9
DeiT [3]		✓	✓	✓	<u>73.2</u>	77.7	**91.9**	92.1
BEiT [1]		✗	✓	✓	71.6	78.6	91.0	**93.9**
BEiT		✗	✗	✓	72.4	<u>79.3</u>	<u>91.7</u>	92.7
SplitMask		✗	✗	✓	**74.6**	**80.4**	91.2	<u>93.1</u>

Robustness w.r.t pre-training dataset

- SplitMask shows a strong transfer performance regardless of the pre-training dataset used. Typically pre-training using the target dataset achieves the strongest result.

Finetuning (→) Pre-training (↓)	iNat-19	iNat-18	Food 101	Cars
Rand Init.	67.5	59.6	84.7	35.3
IMNet	75.8	69.9	**91.5**	92.2
iNat-19	**76.3**	**70.1**	90.4	91.7
iNat-18	75.1	**70.1**	90.4	91.8
Food 101	75.1	68.6	**91.5**	91.7
Cars	71.3	64.2	87.0	92.8
COCO	**76.3**	69.5	90.9	**93.0**

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[3] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.