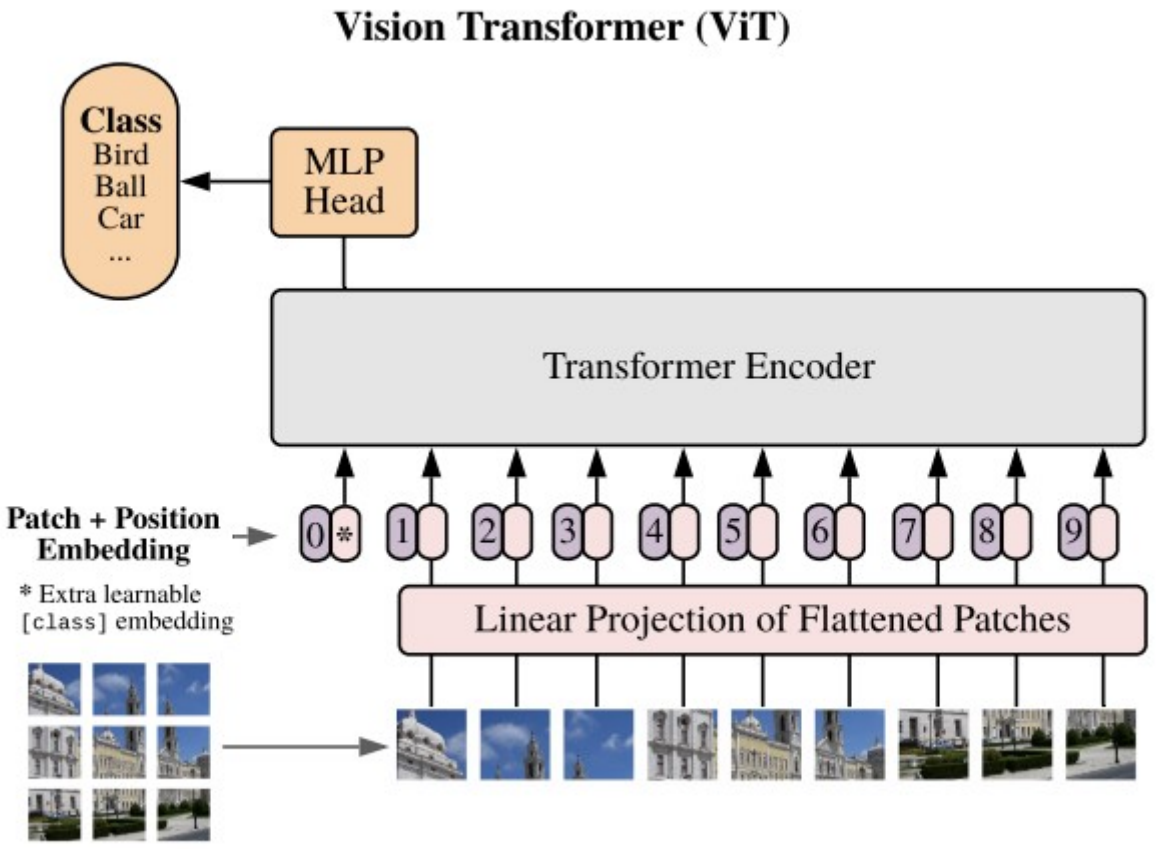
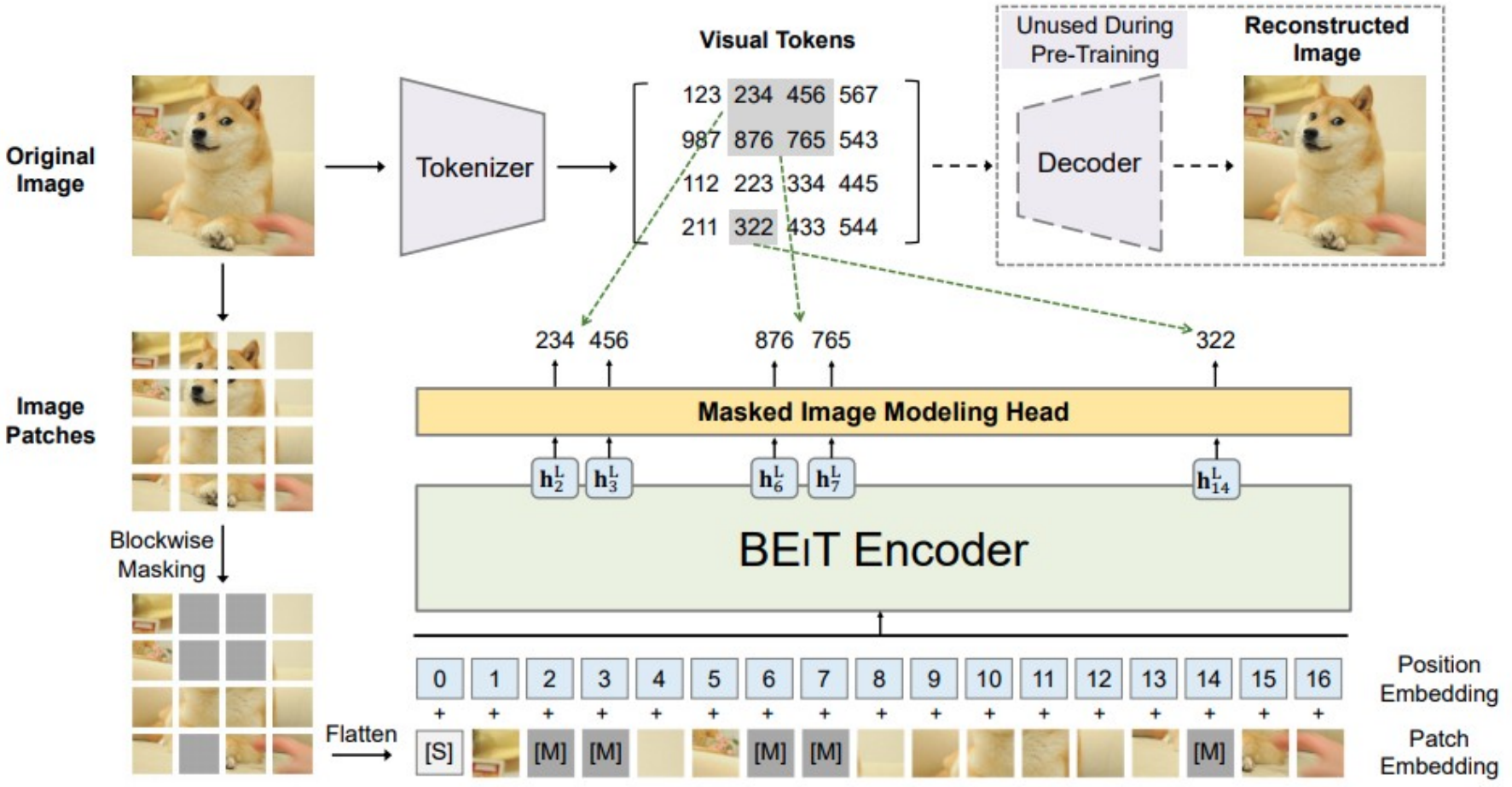
Are Large-scale Datasets Necessary for Self-Supervised Pre-training?

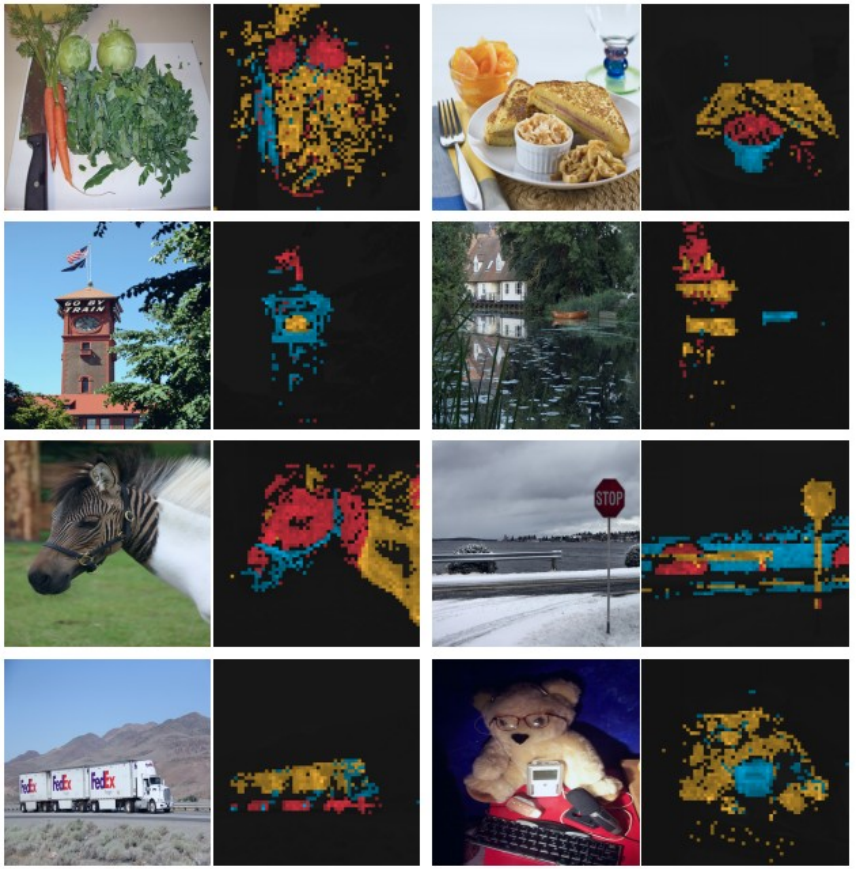
Prerequisites: Vision Transformer

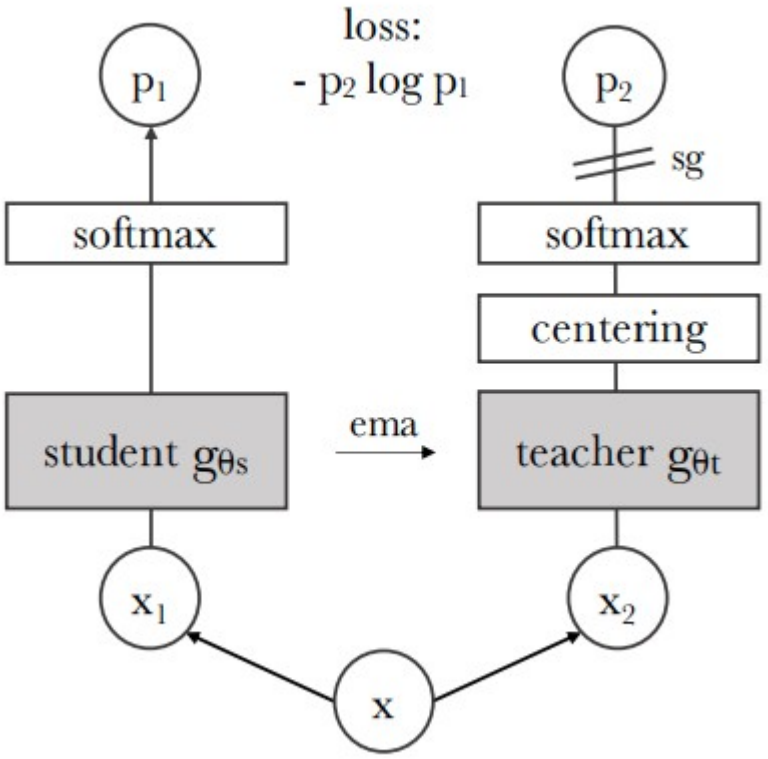


Prerequisites: BEiT

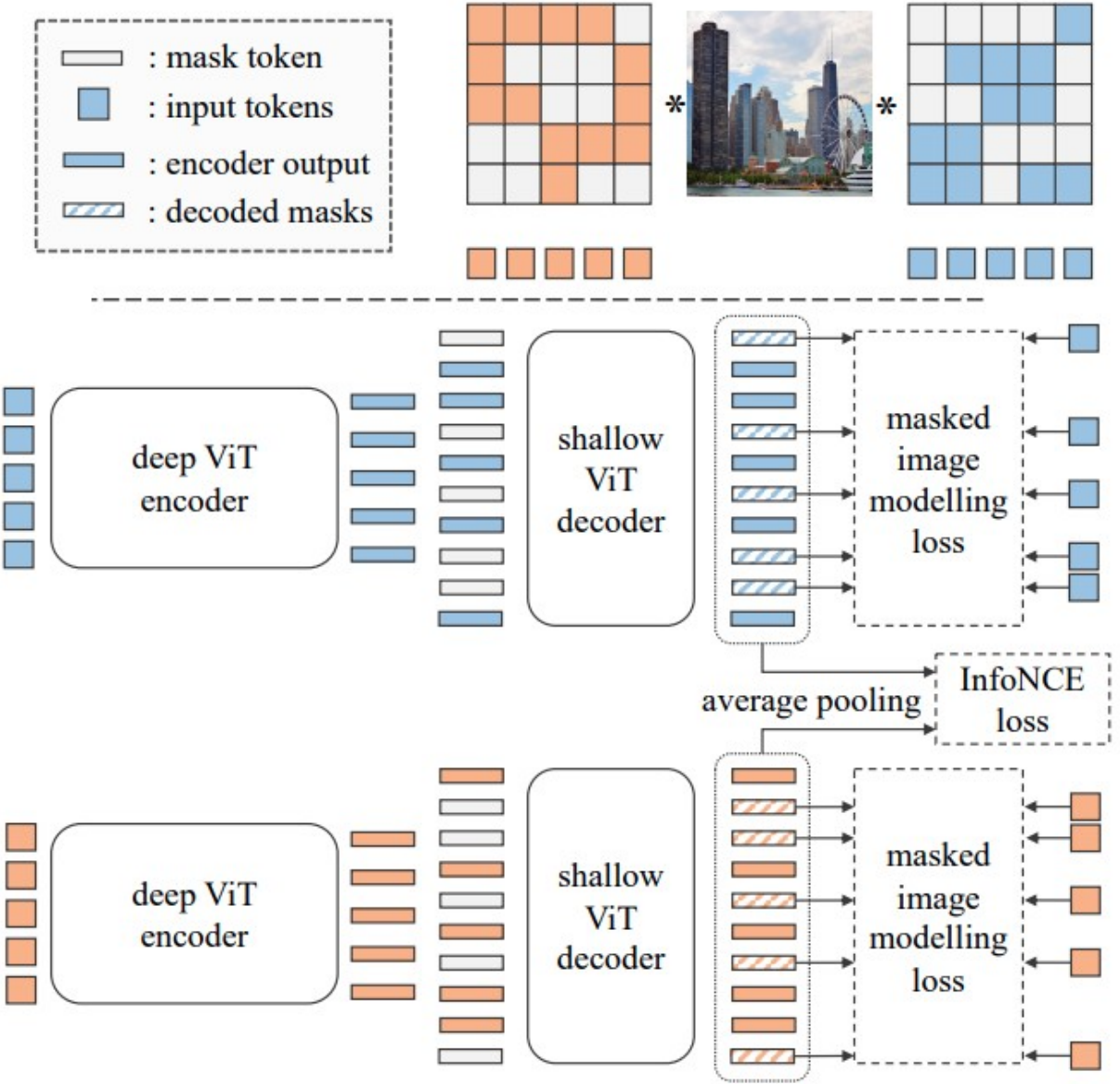


Prerequisites: DINO





SplitMask



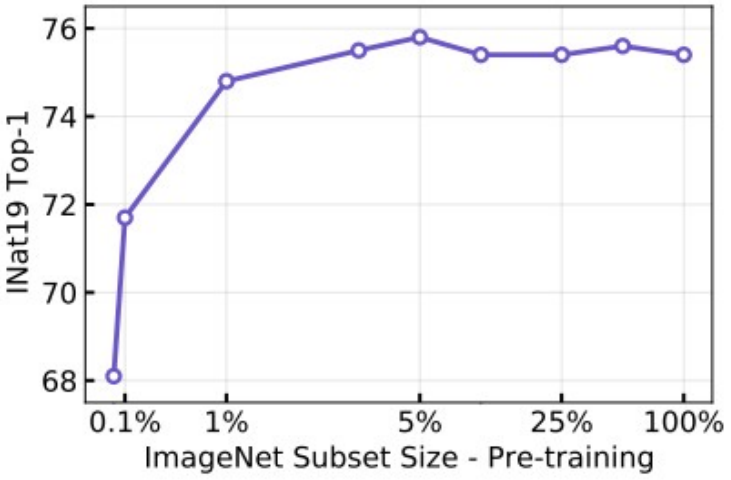
$$\ell(\mathbf{x}_a) = \frac{\exp(\mathbf{x}_a^\top \mathbf{x}_b/\tau)}{\sum_{\mathbf{y}\in\{\mathbf{x}_b\}\cup\mathcal{N}} \exp(\mathbf{x}_a^\top \mathbf{y}/\tau)},$$

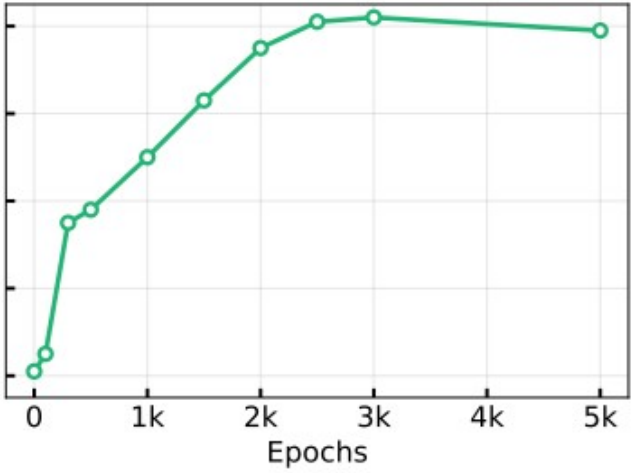
SplitMask: Tokenizers

Table 2. Ablation study on the effect of different tokenization methods. We compare the DALL-E tokenizer originally used in BEiT with patch level techniques: random projection, random patches and k-means clustering. We observe that the DALL-E tokenizer can be effectively replaced by simpler methods that do not require training on a large dataset.

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

SplitMask: Expreriments





Method	IMNet 1% *epochs: 30k*	IMNet 10% *epochs: 3k*	IMNet Full *epochs: 300*	COCO *epochs: 3k*
Supervised	71.6	75.0	75.8	-
DINO [18]	70.1	73.1	**78.4**	71.9
BEiT [24]	74.1	74.5	75.2	74.4
SplitMask	**74.8**	**75.4**	75.4	**76.3**

COCO detection and instance segmentation performance

Method	Backbone	Pre-training			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
		Supervised	IMNet	COCO						
Random Initialization	ViT-S	✗	✗	✗	38.3	60.1	41.4	35.6	57.1	37.7
Random Initialization†		✗	✗	✗	42.8	64.5	45.6	39.1	61.5	41.7
DeiT [50]		✓	✓	✗	44.2	66.6	47.9	40.1	63.2	42.7
BEiT [24]		✗	✓	✗	44.5	66.2	48.8	40.3	63.2	43.1
DINO [18]		✗	✗	✓	43.7	65.5	47.7	39.6	62.3	42.3
BEiT		✗	✗	✓	44.7	66.3	48.8	40.2	63.1	43.2
SplitMask		✗	✗	✓	**45.3**	**66.9**	**49.4**	**40.6**	**63.6**	**43.5**
Random Initialization	ViT-B	✗	✗	✗	40.7	62.7	44.2	37.1	59.1	39.4
Random Initialization†		✗	✗	✗	43.0	64.2	46.9	38.8	61.3	41.6
DeiT [50]		✓	✓	✗	45.5	**67.9**	49.2	41.0	64.6	43.8
BEiT [24]		✗	✓	✗	46.3	67.6	50.6	41.6	64.5	44.9
DINO [18]		✗	✗	✓	43.1	64.4	46.9	38.9	61.4	41.4
BEiT		✗	✗	✓	46.7	67.7	51.2	41.8	65.0	44.6
SplitMask		✗	✗	✓	**46.8**	**67.9**	**51.5**	**42.1**	**65.3**	**45.1**

Pretraining on target dataset

Method	Backbone	Supervised pre-training	Data Used		iNat-18	iNat-19	Food 101	Cars	Clipart	Painting	Sketch
			IMNet	Target	437k	265k	75k	8k	34k	52k	49k
Liu et al. [67]∓	CVT-13	✗	✗	✓	-	-	-	-	60.6	55.2	57.6
	ResNet-50	✗	✗	✓	-	-	-	-	63.9	53.5	59.6
Random Init.	ViT-S	✗	✗	✓	59.6	67.5	84.7	35.3	41.0	38.4	37.2
DeiT [50]		✓	✓	✓	69.9	75.8	**91.5**	92.2	**79.6**	**74.2**	**72.5**
BEiT [24]		✗	✓	✓	68.1	75.2	90.5	92.4	75.3	68.7	68.5
BEiT		✗	✗	✓	68.8	76.1	90.7	92.7	-	69.0	-
SplitMask		✗	✗	✓	**70.1**	**76.3**	**91.5**	**92.8**	78.3	69.2	70.7
Random Init.	ViT-B	✗	✗	✓	59.6	68.1	83.3	36.9	41.9	37.6	34.9
DeiT [50]		✓	✓	✓	73.2	77.7	**91.9**	92.1	**80.0**	**73.8**	**72.6**
BEiT [24]		✗	✓	✓	71.6	78.6	91.0	**93.9**	78.0	71.5	71.4
BEiT		✗	✗	✓	72.4	79.3	91.7	92.7	-	70.7	-
SplitMask		✗	✗	✓	**74.6**	**80.4**	91.2	93.1	79.3	72.0	72.1