XCiT: Cross-Covariance Image Transformers

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek and Hervé Jégou

Presented by Chongyi Zheng and Sabiq Muhtadi

Motivation

Self-attention: Time and memory complexity is **quadratic**

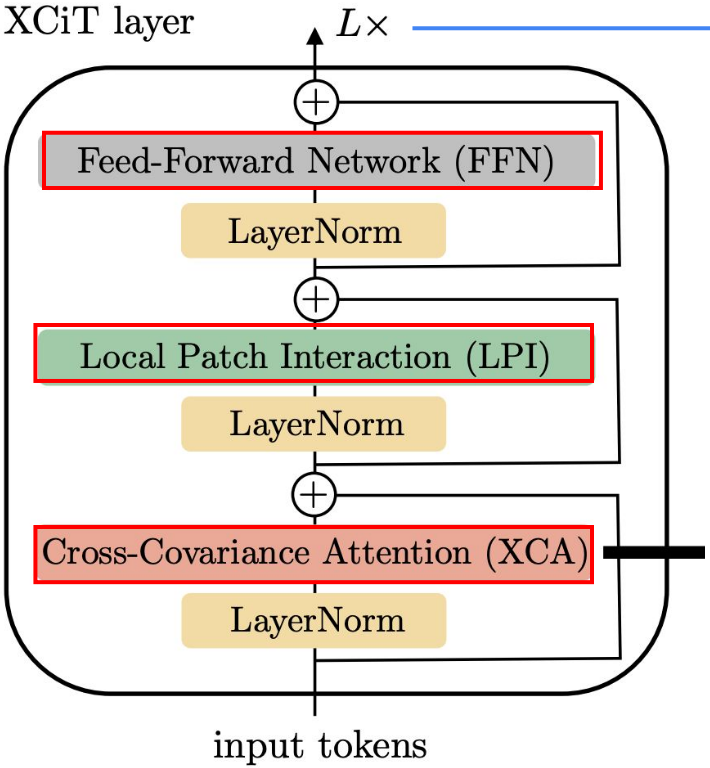
$O(w^2h^2)$ for w×h image

Cross-variance attention: “transposed” self-attention; operates among feature channels, not tokens

Linear to number of patches

XCiT transformer: builds on top of cross-variance attention

Motivation



XCiT blocks for downstream tasks

- Classification
- Dense prediction

Related Work

Deep vision transformers

- model with 48 layers using LayerScale
- residual blocks across layers and improves optimization
- separate patch features and feature aggregation for classification

Spatial structure in vision transformers

- transformer module for intra-patch structure
- LeViT: multi-stage architecture with reduced feature resolution
- convolution-based module for extracting patch descriptors

Related Work

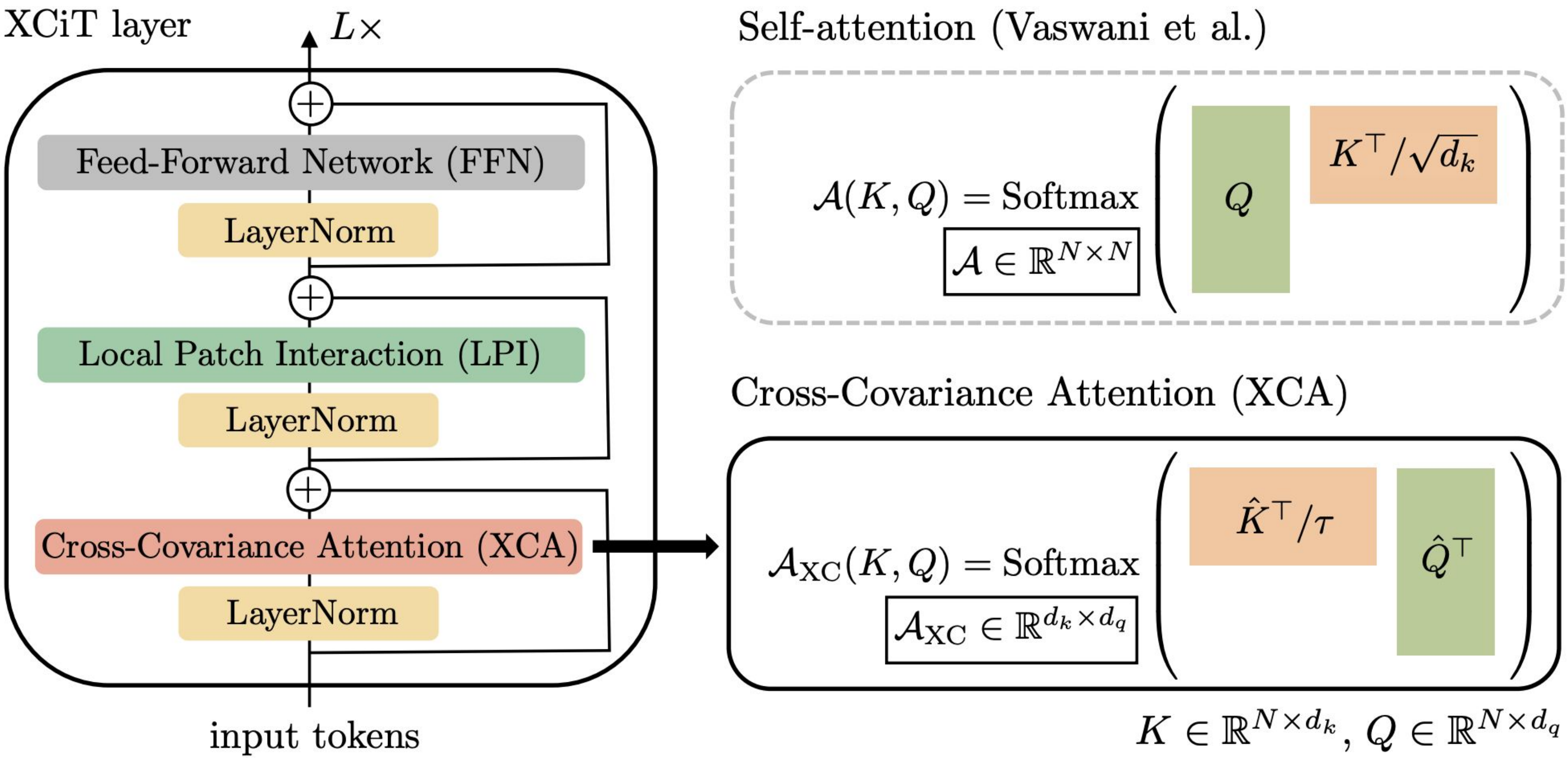
Efficient attention - reduce quadratic complexity

- restrict self-attention to local window, stride, axis
- projection across the token dimension
- factorization of the softmax-attention kernel

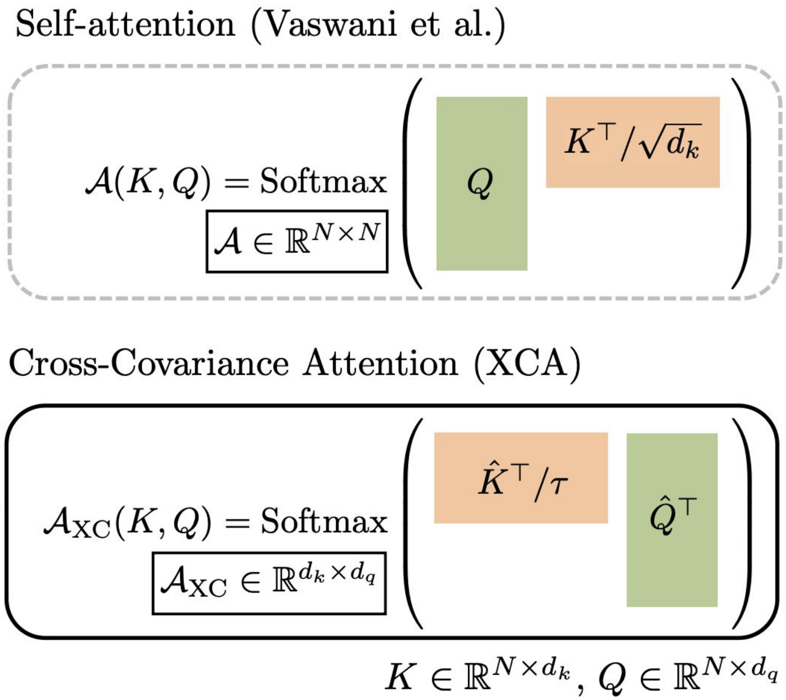
Transformers for high-resolution images.

- pyramidal architecture
- pooling to reduce the resolution across the spatial and temporal dimensions
- global tokens and local attention
- local attention with shifted windows

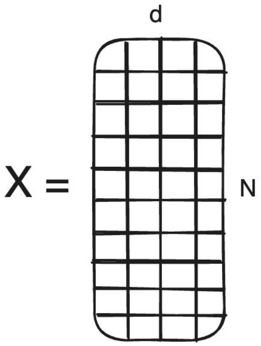
Architecture



Architecture

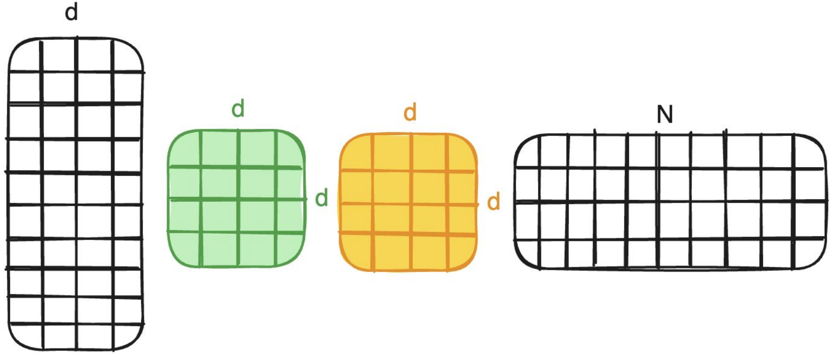


Token self-attention

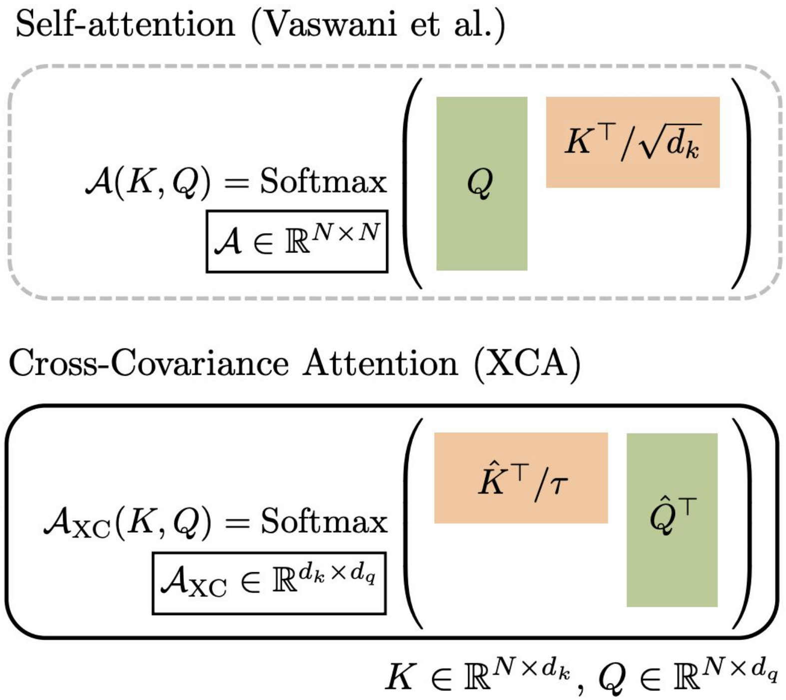


Query Key

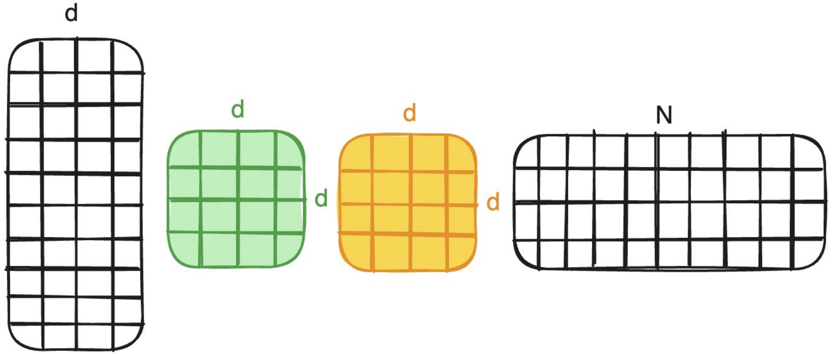
$$XW_q(XW_k)^T = XW_qW_k^TX^T$$



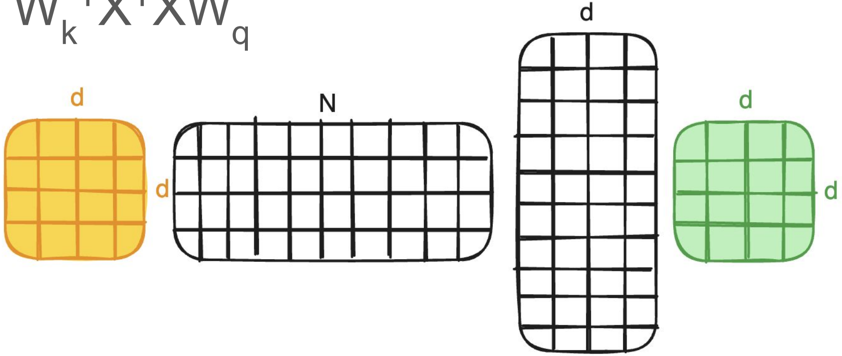
Architecture



$$XW_q(XW_k)^T = XW_qW_k^TX^T$$



$$W_k^TX^TXW_q$$

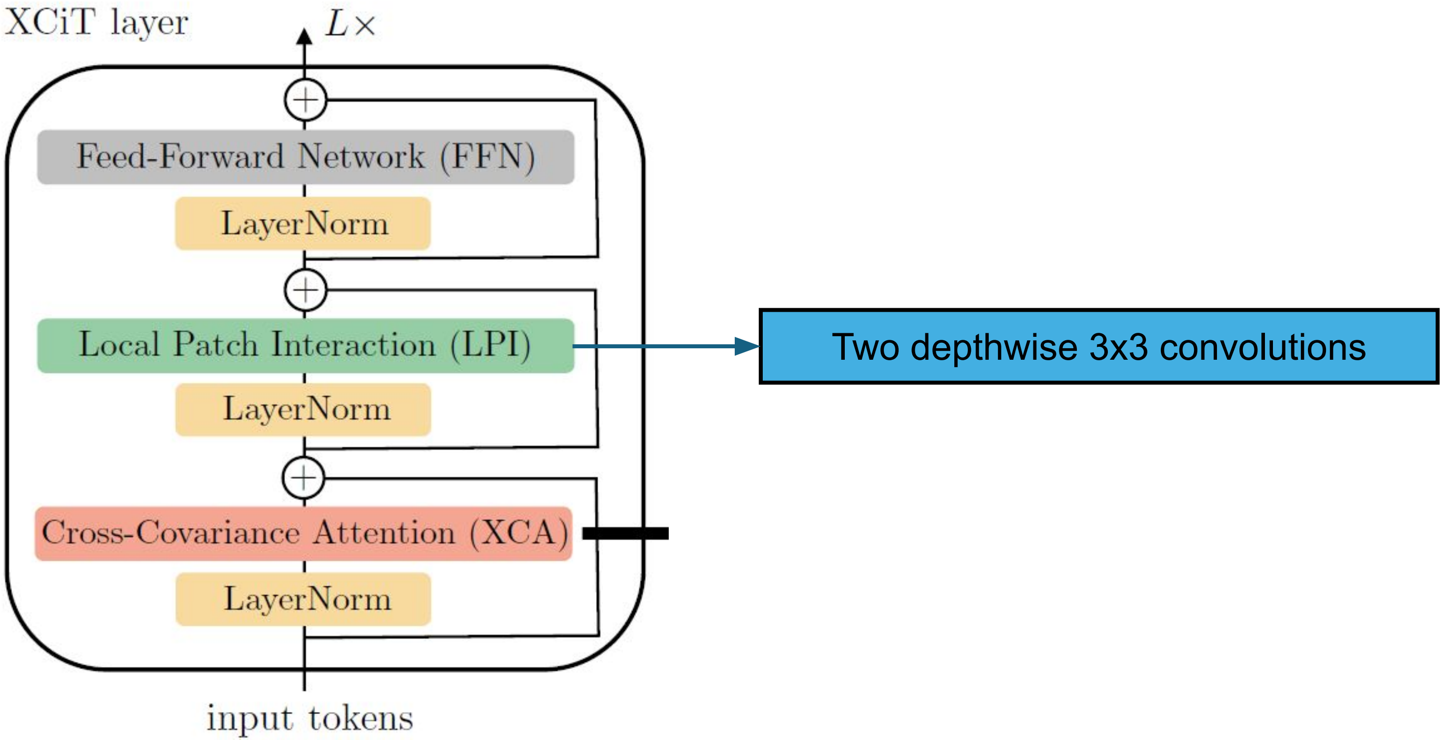


Architecture

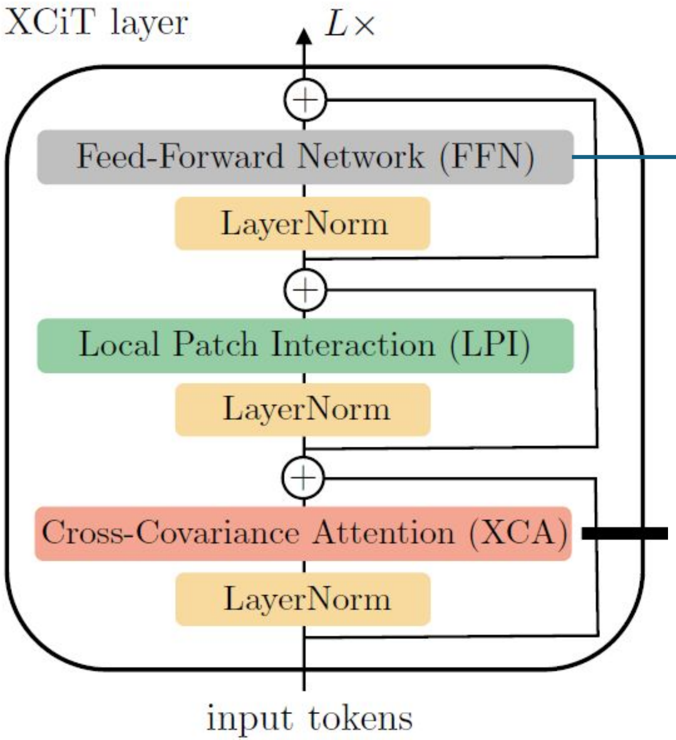
Block-diagonal cross-covariance attention

- divide features into *h* groups
- apply cross-covariance attention separate for each group

Architecture



Architecture



Single Hidden Layer with 4*d* Hidden Units

Image Classification

Table 2: **ImageNet classification**. Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [65].

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-SΥ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24Υ ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16Υ	26M	4.8B	224	83.3	72.5
XCiT-S12/16Υ ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8Υ ↑	26M	55.6B	384	**85.1**	**74.8**
EfficientNet-B7 RA [18]	66M	37.0B	600	84.7	_
NFNet-F0 [10]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [53]	39M	8.0B	224	81.7	72.4
TNT-B [79]	66M	14.1B	224	82.8	_
Swin-S [44]	50M	8.7B	224	83.0	_
CaiT-S24Υ ↑ [68]	47M	32.2B	384	85.1	75.4
XCiT-S24/16Υ	48M	9.1B	224	83.9	73.3
XCiT-S24/16Υ ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8Υ ↑	48M	105.9B	384	**85.6**	**75.7**
Fix-EfficientNet-B8 [66]	87M	89.5B	800	85.7	75.9
RegNetY-16GF [53]	84M	16.0B	224	82.9	72.4
Swin-B↑ [44]	88M	47.0B	384	84.2	_
DeiT-BΥ ↑ [65]	87M	55.5B	384	85.2	75.2
CaiT-S48Υ ↑ [68]	89M	63.8B	384	85.3	**76.2**
XCiT-M24/16Υ	84M	16.2B	224	84.3	73.6
XCiT-M24/16Υ ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8Υ ↑	84M	187.9B	384	**85.8**	76.1
NFNet-F2 [10]	194M	62.6B	352	85.1	74.3
NFNet-F3 [10]	255M	114.8B	416	85.7	75.2
CaiT-M24Υ ↑ [68]	186M	116.1B	384	85.8	76.1
XCiT-L24/16Υ	189M	36.1B	224	84.9	74.6
XCiT-L24/16Υ ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8Υ ↑	189M	417.8B	384	**86.0**	**76.6**

Image Classification

Table 2: **ImageNet classification**. Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [65].

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-Sϒ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24ϒ ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16ϒ	26M	4.8B	224	83.3	72.5
XCiT-S12/16ϒ ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8ϒ ↑	26M	55.6B	384	**85.1**	**74.8**
EfficientNet-B7 RA [18]	66M	37.0B	600	84.7	_
NFNet-F0 [10]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [53]	39M	8.0B	224	81.7	72.4
TNT-B [79]	66M	14.1B	224	82.8	_
Swin-S [44]	50M	8.7B	224	83.0	_
CaiT-S24ϒ ↑ [68]	47M	32.2B	384	85.1	75.4
XCiT-S24/16ϒ	48M	9.1B	224	83.9	73.3
XCiT-S24/16ϒ ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8ϒ ↑	48M	105.9B	384	**85.6**	**75.7**
Fix-EfficientNet-B8 [66]	87M	89.5B	800	85.7	75.9
RegNetY-16GF [53]	84M	16.0B	224	82.9	72.4
Swin-B↑ [44]	88M	47.0B	384	84.2	_
DeiT-Bϒ ↑ [65]	87M	55.5B	384	85.2	75.2
CaiT-S48ϒ ↑ [68]	89M	63.8B	384	85.3	**76.2**
XCiT-M24/16ϒ	84M	16.2B	224	84.3	73.6
XCiT-M24/16ϒ ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8ϒ ↑	84M	187.9B	384	**85.8**	76.1
NFNet-F2 [10]	194M	62.6B	352	85.1	74.3
NFNet-F3 [10]	255M	114.8B	416	85.7	75.2
CaiT-M24ϒ ↑ [68]	186M	116.1B	384	85.8	76.1
XCiT-L24/16ϒ	189M	36.1B	224	84.9	74.6
XCiT-L24/16ϒ ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8ϒ ↑	189M	417.8B	384	**86.0**	**76.6**

Image Classification

Table 2: **ImageNet classification**. Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [65].

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-Sϒ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24ϒ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16ϒ	26M	4.8B	224	83.3	72.5
XCiT-S12/16ϒ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8ϒ↑	26M	55.6B	384	**85.1**	**74.8**
EfficientNet-B7 RA [18]	66M	37.0B	600	84.7	_
NFNet-F0 [10]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [53]	39M	8.0B	224	81.7	72.4
TNT-B [79]	66M	14.1B	224	82.8	_
Swin-S [44]	50M	8.7B	224	83.0	_
CaiT-S24ϒ ↑ [68]	47M	32.2B	384	85.1	75.4
XCiT-S24/16ϒ	48M	9.1B	224	83.9	73.3
XCiT-S24/16ϒ ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8ϒ ↑	48M	105.9B	384	**85.6**	**75.7**
Fix-EfficientNet-B8 [66]	87M	89.5B	800	85.7	75.9
RegNetY-16GF [53]	84M	16.0B	224	82.9	72.4
Swin-B↑ [44]	88M	47.0B	384	84.2	_
DeiT-Bϒ ↑ [65]	87M	55.5B	384	85.2	75.2
CaiT-S48ϒ ↑ [68]	89M	63.8B	384	85.3	**76.2**
XCiT-M24/16ϒ	84M	16.2B	224	84.3	73.6
XCiT-M24/16ϒ ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8ϒ ↑	84M	187.9B	384	**85.8**	76.1
NFNet-F2 [10]	194M	62.6B	352	85.1	74.3
NFNet-F3 [10]	255M	114.8B	416	85.7	75.2
CaiT-M24ϒ ↑ [68]	186M	116.1B	384	85.8	76.1
XCiT-L24/16ϒ	189M	36.1B	224	84.9	74.6
XCiT-L24/16ϒ ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8ϒ ↑	189M	417.8B	384	**86.0**	**76.6**

Image Classification

Table 2: **ImageNet classification**. Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [65].

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-SΥ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24Υ ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16Υ	26M	4.8B	224	83.3	72.5
XCiT-S12/16Υ ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8Υ ↑	26M	55.6B	384	**85.1**	**74.8**
EfficientNet-B7 RA [18]	66M	37.0B	600	84.7	_
NFNet-F0 [10]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [53]	39M	8.0B	224	81.7	72.4
TNT-B [79]	66M	14.1B	224	82.8	_
Swin-S [44]	50M	8.7B	224	83.0	_
CaiT-S24Υ ↑ [68]	47M	32.2B	384	85.1	75.4
XCiT-S24/16Υ	48M	9.1B	224	83.9	73.3
XCiT-S24/16Υ ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8Υ ↑	48M	105.9B	384	**85.6**	**75.7**
Fix-EfficientNet-B8 [66]	87M	89.5B	800	85.7	75.9
RegNetY-16GF [53]	84M	16.0B	224	82.9	72.4
Swin-B↑ [44]	88M	47.0B	384	84.2	_
DeiT-BΥ ↑ [65]	87M	55.5B	384	85.2	75.2
CaiT-S48Υ ↑ [68]	89M	63.8B	384	85.3	**76.2**
XCiT-M24/16Υ	84M	16.2B	224	84.3	73.6
XCiT-M24/16Υ ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8Υ ↑	84M	187.9B	384	**85.8**	76.1
NFNet-F2 [10]	194M	62.6B	352	85.1	74.3
NFNet-F3 [10]	255M	114.8B	416	85.7	75.2
CaiT-M24Υ ↑ [68]	186M	116.1B	384	85.8	76.1
XCiT-L24/16Υ	189M	36.1B	224	84.9	74.6
XCiT-L24/16Υ ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8Υ ↑	189M	417.8B	384	**86.0**	**76.6**

Image Classification

Table 2: **ImageNet classification**. Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [65].

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-SΥ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24Υ ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16Υ	26M	4.8B	224	83.3	72.5
XCiT-S12/16Υ ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8Υ ↑	26M	55.6B	384	**85.1**	**74.8**
EfficientNet-B7 RA [18]	66M	37.0B	600	84.7	_
NFNet-F0 [10]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [53]	39M	8.0B	224	81.7	72.4
TNT-B [79]	66M	14.1B	224	82.8	_
Swin-S [44]	50M	8.7B	224	83.0	_
CaiT-S24Υ ↑ [68]	47M	32.2B	384	85.1	75.4
XCiT-S24/16Υ	48M	9.1B	224	83.9	73.3
XCiT-S24/16Υ ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8Υ ↑	48M	105.9B	384	**85.6**	**75.7**
Fix-EfficientNet-B8 [66]	87M	89.5B	800	85.7	75.9
RegNetY-16GF [53]	84M	16.0B	224	82.9	72.4
Swin-B↑ [44]	88M	47.0B	384	84.2	_
DeiT-BΥ ↑ [65]	87M	55.5B	384	85.2	75.2
CaiT-S48Υ ↑ [68]	89M	63.8B	384	85.3	**76.2**
XCiT-M24/16Υ	84M	16.2B	224	84.3	73.6
XCiT-M24/16Υ ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8Υ ↑	84M	187.9B	384	**85.8**	76.1
NFNet-F2 [10]	194M	62.6B	352	85.1	74.3
NFNet-F3 [10]	255M	114.8B	416	85.7	75.2
CaiT-M24Υ ↑ [68]	186M	116.1B	384	85.8	76.1
XCiT-L24/16Υ	189M	36.1B	224	84.9	74.6
XCiT-L24/16Υ ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8Υ ↑	189M	417.8B	384	**86.0**	**76.6**

Image Classification

Table 2: **ImageNet classification**. Number of parameters, FLOPs, image resolution, and top-1 accuracy on ImageNet-1k and ImageNet-V2. Training strategies vary across models, transformer-based models and the reported RegNet mostly follow recipes from DeiT [65].

Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-Sϒ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24ϒ ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16ϒ	26M	4.8B	224	83.3	72.5
XCiT-S12/16ϒ ↑	26M	14.3B	384	84.7	74.1
XCiT-S12/8ϒ ↑	26M	55.6B	384	**85.1**	**74.8**
EfficientNet-B7 RA [18]	66M	37.0B	600	84.7	_
NFNet-F0 [10]	72M	12.4B	256	83.6	72.6
RegNetY-8GF [53]	39M	8.0B	224	81.7	72.4
TNT-B [79]	66M	14.1B	224	82.8	_
Swin-S [44]	50M	8.7B	224	83.0	_
CaiT-S24ϒ ↑ [68]	47M	32.2B	384	85.1	75.4
XCiT-S24/16ϒ	48M	9.1B	224	83.9	73.3
XCiT-S24/16ϒ ↑	48M	26.9B	384	85.1	74.6
XCiT-S24/8ϒ ↑	48M	105.9B	384	**85.6**	**75.7**
Fix-EfficientNet-B8 [66]	87M	89.5B	800	85.7	75.9
RegNetY-16GF [53]	84M	16.0B	224	82.9	72.4
Swin-B↑ [44]	88M	47.0B	384	84.2	_
DeiT-Bϒ ↑ [65]	87M	55.5B	384	85.2	75.2
CaiT-S48ϒ ↑ [68]	89M	63.8B	384	85.3	**76.2**
XCiT-M24/16ϒ	84M	16.2B	224	84.3	73.6
XCiT-M24/16ϒ ↑	84M	47.7B	384	85.4	75.1
XCiT-M24/8ϒ ↑	84M	187.9B	384	**85.8**	76.1
NFNet-F2 [10]	194M	62.6B	352	85.1	74.3
NFNet-F3 [10]	255M	114.8B	416	85.7	75.2
CaiT-M24ϒ ↑ [68]	186M	116.1B	384	85.8	76.1
XCiT-L24/16ϒ	189M	36.1B	224	84.9	74.6
XCiT-L24/16ϒ ↑	189M	106.0B	384	85.8	75.8
XCiT-L24/8ϒ ↑	189M	417.8B	384	**86.0**	**76.6**

Image Classification

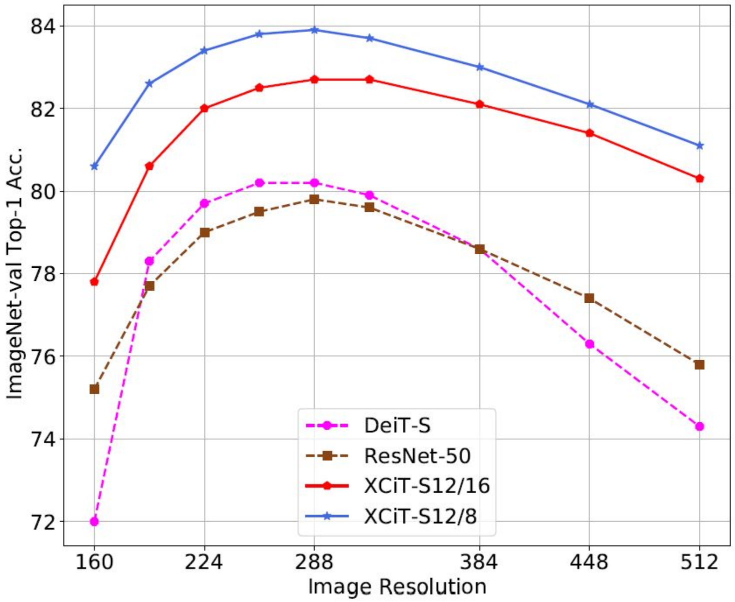


Figure 3: Performance when changing the resolution at test-time for models with a similar number of parameters. All networks were trained at resolution 224, w/o distillation. XCiT is more tolerant to changes of resolution than the Gram-based DeiT and benefit more from the "FixRes" effect [64] when inference is performed at a larger resolution than at train-time.

Image Classification – Self Supervised Learning

Table 3: **Self-supervised learning.** Top-1 acc. on ImageNet-1k. Wwe report with a crop-ratio 0.875 for consistency with DINO. For the last row it is set to 1.0 (improves from 80.7% to 80.9%). All models are trained for 300 epochs.

SSL Method	Model	#params	FLOPs	Linear	k-NN
MoBY [76]	Swin-T [44]	29M	4.5B	75.0	–
DINO [12]	ResNet-50 [28]	23M	4.1B	74.5	65.6
DINO [12]	ViT-S/16 [22]	22M	4.6B	76.1	72.8
DINO [12]	ViT-S/8 [22]	22M	22.4B	**79.2**	**77.2**
DINO [12]	XCiT-S12/16	26M	4.9B	77.8	76.0
DINO [12]	XCiT-S12/8	26M	18.9B	**79.2**	77.1
DINO [12]	ViT-B/16 [22]	87M	17.5B	78.2	76.1
DINO [12]	ViT-B/8 [22]	87M	78.2B	80.1	77.4
DINO [12]	XCiT-M24/16	84M	16.2B	78.8	76.4
DINO [12]	XCiT-M24/8	84M	64.0B	**80.3**	**77.9**
DINO [12]	XCiT-M24/8↑384	84M	188.0B	**80.9**	-

Image Classification – Self Supervised Learning

Table 3: **Self-supervised learning.** Top-1 acc. on ImageNet-1k. Wwe report with a crop-ratio 0.875 for consistency with DINO. For the last row it is set to 1.0 (improves from 80.7% to 80.9%). All models are trained for 300 epochs.

SSL Method	Model	#params	FLOPs	Linear	k-NN
MoBY [76]	Swin-T [44]	29M	4.5B	75.0	–
DINO [12]	ResNet-50 [28]	23M	4.1B	74.5	65.6
DINO [12]	ViT-S/16 [22]	22M	4.6B	76.1	72.8
DINO [12]	ViT-S/8 [22]	22M	22.4B	**79.2**	**77.2**
DINO [12]	XCiT-S12/16	26M	4.9B	77.8	76.0
DINO [12]	XCiT-S12/8	26M	18.9B	**79.2**	77.1
DINO [12]	ViT-B/16 [22]	87M	17.5B	78.2	76.1
DINO [12]	ViT-B/8 [22]	87M	78.2B	80.1	77.4
DINO [12]	XCiT-M24/16	84M	16.2B	78.8	76.4
DINO [12]	XCiT-M24/8	84M	64.0B	**80.3**	**77.9**
DINO [12]	XCiT-M24/8↑384	84M	188.0B	**80.9**	-

Image Classification – Self Supervised Learning

Table 3: **Self-supervised learning.** Top-1 acc. on ImageNet-1k. Wwe report with a crop-ratio 0.875 for consistency with DINO. For the last row it is set to 1.0 (improves from 80.7% to 80.9%). All models are trained for 300 epochs.

SSL Method	Model	#params	FLOPs	Linear	k-NN
MoBY [76]	Swin-T [44]	29M	4.5B	75.0	–
DINO [12]	ResNet-50 [28]	23M	4.1B	74.5	65.6
DINO [12]	ViT-S/16 [22]	22M	4.6B	76.1	72.8
DINO [12]	ViT-S/8 [22]	22M	22.4B	**79.2**	**77.2**
DINO [12]	XCiT-S12/16	26M	4.9B	77.8	76.0
DINO [12]	XCiT-S12/8	26M	18.9B	**79.2**	77.1
DINO [12]	ViT-B/16 [22]	87M	17.5B	78.2	76.1
DINO [12]	ViT-B/8 [22]	87M	78.2B	80.1	77.4
DINO [12]	XCiT-M24/16	84M	16.2B	78.8	76.4
DINO [12]	XCiT-M24/8	84M	64.0B	**80.3**	**77.9**
DINO [12]	XCiT-M24/8↑384	84M	188.0B	**80.9**	-

Image Classification - Ablations

Table 4: **Ablations** of various architectural design choices on the task of ImageNet-1k classification using the XCiT-S12 model. Our baseline model uses the convolutional projection adopted from LeVit.

Model	Ablation	ImNet top-1 acc.
XCiT-S12/16	Baseline	82.0
XCiT-S12/8		83.4
XCiT-S12/16	Linear patch proj.	81.1
XCiT-S12/8		83.1
XCiT-S12/16	w/o LPI layer	80.8
	w/o XCA layer	75.9
XCiT-S12/16	w/o ℓ_2-normal.	failed
	w/o learned temp. τ	81.8

Image Classification - Ablations

Table 4: **Ablations** of various architectural design choices on the task of ImageNet-1k classification using the XCiT-S12 model. Our baseline model uses the convolutional projection adopted from LeVit.

Model	Ablation	ImNet top-1 acc.
XCiT-S12/16	Baseline	82.0
XCiT-S12/8		83.4
XCiT-S12/16	Linear patch proj.	81.1
XCiT-S12/8		83.1
XCiT-S12/16	w/o LPI layer	80.8
	w/o XCA layer	75.9
XCiT-S12/16	w/o ℓ_2-normal.	failed
	w/o learned temp. τ	81.8

Image Classification - Ablations

Table 4: **Ablations** of various architectural design choices on the task of ImageNet-1k classification using the XCiT-S12 model. Our baseline model uses the convolutional projection adopted from LeVit.

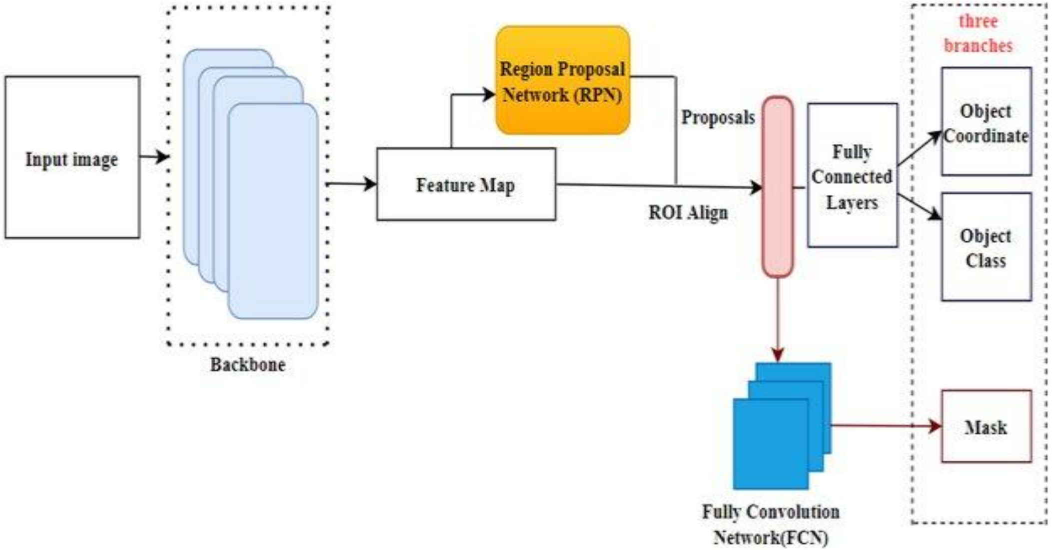
Model	Ablation	ImNet top-1 acc.
XCiT-S12/16	Baseline	82.0
XCiT-S12/8		83.4
XCiT-S12/16	Linear patch proj.	81.1
XCiT-S12/8		83.1
XCiT-S12/16	w/o LPI layer	80.8
	w/o XCA layer	75.9
XCiT-S12/16	w/o ℓ_2-normal.	failed
	w/o learned temp. τ	81.8

Image Classification - Ablations

Table 4: **Ablations** of various architectural design choices on the task of ImageNet-1k classification using the XCiT-S12 model. Our baseline model uses the convolutional projection adopted from LeVit.

Model	Ablation	ImNet top-1 acc.
XCiT-S12/16	Baseline	82.0
XCiT-S12/8		83.4
XCiT-S12/16	Linear patch proj.	81.1
XCiT-S12/8		83.1
XCiT-S12/16	w/o LPI layer	80.8
	w/o XCA layer	75.9
XCiT-S12/16	w/o ℓ_2-normal.	failed
	w/o learned temp. τ	81.8

Object Detection and Instance Segmentation



Object Detection and Instance Segmentation

Table 5: **COCO object detection and instance segmentation** performance on the mini-val set. All backbones are pre-trained on ImageNet-1k, use Mask R-CNN model [29] and are trained with the same 3x schedule.

Backbone	#params	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet18 [28]	31.2M	36.9	57.1	40.0	33.6	53.9	35.7
PVT-Tiny [71]	32.9M	39.8	62.2	43.0	37.4	59.3	39.9
ViL-Tiny [81]	26.9M	41.2	64.0	44.7	37.9	59.8	40.6
XCiT-T12/16	26.1M	42.7	64.3	46.4	38.5	61.2	41.1
XCiT-T12/8	25.8M	**44.5**	**66.4**	**48.8**	**40.3**	**63.5**	**43.2**
ResNet50 [28]	44.2M	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small [71]	44.1M	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [81]	45.0M	43.4	64.9	47.0	39.6	62.1	42.4
Swin-T [44]	47.8M	46.0	68.1	50.3	41.6	65.1	44.9
XCiT-S12/16	44.3M	45.3	67.0	49.5	40.8	64.0	43.8
XCiT-S12/8	43.1M	**47.0**	**68.9**	**51.7**	**42.3**	**66.0**	**45.4**
ResNet101 [28]	63.2M	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32	62.8M	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [71]	63.9M	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [81]	60.1M	44.6	66.3	48.5	40.7	63.8	43.7
Swin-S [44]	69.1M	**48.5**	**70.2**	**53.5**	**43.3**	**67.3**	**46.6**
XCiT-S24/16	65.8M	46.5	68.0	50.9	41.8	65.2	45.0
XCiT-S24/8	64.5M	48.1	69.5	53.0	43.0	66.5	46.1
ResNeXt101-64 [75]	101.9M	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large [71]	81.0M	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Large [81]	76.1M	45.7	67.2	49.9	41.3	64.4	44.5
XCiT-M24/16	101.1M	46.7	68.2	51.1	42.0	65.6	44.9
XCiT-M24/8	98.9M	**48.5**	**70.3**	**53.4**	**43.7**	**67.5**	**46.9**

Object Detection and Instance Segmentation

Table 5: **COCO object detection and instance segmentation** performance on the mini-val set. All backbones are pre-trained on ImageNet-1k, use Mask R-CNN model [29] and are trained with the same 3x schedule.

Backbone	#params	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet18 [28]	31.2M	36.9	57.1	40.0	33.6	53.9	35.7
PVT-Tiny [71]	32.9M	39.8	62.2	43.0	37.4	59.3	39.9
ViL-Tiny [81]	26.9M	41.2	64.0	44.7	37.9	59.8	40.6
XCiT-T12/16	26.1M	42.7	64.3	46.4	38.5	61.2	41.1
XCiT-T12/8	25.8M	**44.5**	**66.4**	**48.8**	**40.3**	**63.5**	**43.2**
ResNet50 [28]	44.2M	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small [71]	44.1M	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [81]	45.0M	43.4	64.9	47.0	39.6	62.1	42.4
Swin-T [44]	47.8M	46.0	68.1	50.3	41.6	65.1	44.9
XCiT-S12/16	44.3M	45.3	67.0	49.5	40.8	64.0	43.8
XCiT-S12/8	43.1M	**47.0**	**68.9**	**51.7**	**42.3**	**66.0**	**45.4**
ResNet101 [28]	63.2M	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32	62.8M	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [71]	63.9M	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [81]	60.1M	44.6	66.3	48.5	40.7	63.8	43.7
Swin-S [44]	69.1M	**48.5**	**70.2**	**53.5**	**43.3**	**67.3**	**46.6**
XCiT-S24/16	65.8M	46.5	68.0	50.9	41.8	65.2	45.0
XCiT-S24/8	64.5M	48.1	69.5	53.0	43.0	66.5	46.1
ResNeXt101-64 [75]	101.9M	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large [71]	81.0M	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Large [81]	76.1M	45.7	67.2	49.9	41.3	64.4	44.5
XCiT-M24/16	101.1M	46.7	68.2	51.1	42.0	65.6	44.9
XCiT-M24/8	98.9M	**48.5**	**70.3**	**53.4**	**43.7**	**67.5**	**46.9**

Object Detection and Instance Segmentation

Table 5: **COCO object detection and instance segmentation** performance on the mini-val set. All backbones are pre-trained on ImageNet-1k, use Mask R-CNN model [29] and are trained with the same 3x schedule.

Backbone	#params	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet18 [28]	31.2M	36.9	57.1	40.0	33.6	53.9	35.7
PVT-Tiny [71]	32.9M	39.8	62.2	43.0	37.4	59.3	39.9
ViL-Tiny [81]	26.9M	41.2	64.0	44.7	37.9	59.8	40.6
XCiT-T12/16	26.1M	42.7	64.3	46.4	38.5	61.2	41.1
XCiT-T12/8	25.8M	**44.5**	**66.4**	**48.8**	**40.3**	**63.5**	**43.2**
ResNet50 [28]	44.2M	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small [71]	44.1M	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [81]	45.0M	43.4	64.9	47.0	39.6	62.1	42.4
Swin-T [44]	47.8M	46.0	68.1	50.3	41.6	65.1	44.9
XCiT-S12/16	44.3M	45.3	67.0	49.5	40.8	64.0	43.8
XCiT-S12/8	43.1M	**47.0**	**68.9**	**51.7**	**42.3**	**66.0**	**45.4**
ResNet101 [28]	63.2M	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32	62.8M	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [71]	63.9M	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [81]	60.1M	44.6	66.3	48.5	40.7	63.8	43.7
Swin-S [44]	69.1M	**48.5**	**70.2**	**53.5**	**43.3**	**67.3**	**46.6**
XCiT-S24/16	65.8M	46.5	68.0	50.9	41.8	65.2	45.0
XCiT-S24/8	64.5M	48.1	69.5	53.0	43.0	66.5	46.1
ResNeXt101-64 [75]	101.9M	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large [71]	81.0M	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Large [81]	76.1M	45.7	67.2	49.9	41.3	64.4	44.5
XCiT-M24/16	101.1M	46.7	68.2	51.1	42.0	65.6	44.9
XCiT-M24/8	98.9M	**48.5**	**70.3**	**53.4**	**43.7**	**67.5**	**46.9**

Object Detection and Instance Segmentation

Table 5: **COCO object detection and instance segmentation** performance on the mini-val set. All backbones are pre-trained on ImageNet-1k, use Mask R-CNN model [29] and are trained with the same 3x schedule.

Backbone	#params	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet18 [28]	31.2M	36.9	57.1	40.0	33.6	53.9	35.7
PVT-Tiny [71]	32.9M	39.8	62.2	43.0	37.4	59.3	39.9
ViL-Tiny [81]	26.9M	41.2	64.0	44.7	37.9	59.8	40.6
XCiT-T12/16	26.1M	42.7	64.3	46.4	38.5	61.2	41.1
XCiT-T12/8	25.8M	**44.5**	**66.4**	**48.8**	**40.3**	**63.5**	**43.2**
ResNet50 [28]	44.2M	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small [71]	44.1M	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [81]	45.0M	43.4	64.9	47.0	39.6	62.1	42.4
Swin-T [44]	47.8M	46.0	68.1	50.3	41.6	65.1	44.9
XCiT-S12/16	44.3M	45.3	67.0	49.5	40.8	64.0	43.8
XCiT-S12/8	43.1M	**47.0**	**68.9**	**51.7**	**42.3**	**66.0**	**45.4**
ResNet101 [28]	63.2M	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32	62.8M	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [71]	63.9M	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [81]	60.1M	44.6	66.3	48.5	40.7	63.8	43.7
Swin-S [44]	69.1M	**48.5**	**70.2**	**53.5**	**43.3**	**67.3**	**46.6**
XCiT-S24/16	65.8M	46.5	68.0	50.9	41.8	65.2	45.0
XCiT-S24/8	64.5M	48.1	69.5	53.0	43.0	66.5	46.1
ResNeXt101-64 [75]	101.9M	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large [71]	81.0M	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Large [81]	76.1M	45.7	67.2	49.9	41.3	64.4	44.5
XCiT-M24/16	101.1M	46.7	68.2	51.1	42.0	65.6	44.9
XCiT-M24/8	98.9M	**48.5**	**70.3**	**53.4**	**43.7**	**67.5**	**46.9**

Semantic Segmentation

Table 6: **ADE20k semantic segmentation** performance using Semantic FPN [38] and UperNet [74] (in comparable settings). We do not include comparisons with other state-of-the-art models that are pre-trained on larger datasets [44, 54, 83].

Backbone	Semantic FPN		UperNet	
	#params	mIoU	#params	mIoU
ResNet18 [28]	15.5M	32.9	-	-
PVT-Tiny [71]	17.0M	35.7M	-	-
XCiT-T12/16	8.4M	38.1	33.7M	41.5
XCiT-T12/8	8.4M	**39.9**	33.7	**43.5**
ResNet50 [28]	28.5M	36.7	66.5M	42.0
PVT-Small [71]	28.2M	39.8	-	-
Swin-T [44]	-	-	59.9M	44.5
XCiT-S12/16	30.4M	43.9	52.4M	45.9
XCiT-S12/8	30.4M	**44.2**	52.3M	**46.6**
ResNet101 [28]	47.5M	38.8	85.5M	43.8
ResNeXt101-32 [75]	47.1M	39.7	-	-
PVT-Medium [71]	48.0M	41.6	-	-
Swin-S [44]	-	-	81.0M	47.6
XCiT-S24/16	51.8M	44.6	73.8M	46.9
XCiT-S24/8	51.8M	**47.1**	73.8M	**48.1**
ResNeXt101-64 [75]	86.4M	40.2	-	-
PVT-Large [71]	65.1M	42.1	-	-
Swin-B [44]	-	-	121.0M	48.1
XCiT-M24/16	90.8M	45.9	109.0M	47.6
XCiT-M24/8	90.8M	**46.9**	108.9M	**48.4**

Conclusion

- XCiT Models are built using **cross-covariance attention** as the core component
  - Alternative to token self attention that **operates on the feature dimension**
  - **Eliminates** need for **expensive computation** of quadratic attention maps

Conclusion

- XCiT Models are built using **cross-covariance attention** as the core component
  - Alternative to token self attention that **operates on the feature dimension**
  - **Eliminates** need for **expensive computation** of quadratic attention maps

- XCiT exhibits **strong image classification performance**
  - On par with state-of-the-art transformer models
  - **Robust to changing image resolutions**

Conclusion

- XCiT Models are built using **cross-covariance attention** as the core component
  - Alternative to token self attention that **operates on the feature dimension**
  - **Eliminates** need for **expensive computation** of quadratic attention maps

- XCiT exhibits **strong image classification performance**
  - On par with state-of-the-art transformer models
  - **Robust to changing image resolutions**

- XCiT acts as a **very effective backbone** for dense prediction tasks such as **object detection**, **instance and semantic segmentation**

Conclusion

- XCiT Models are built using **cross-covariance attention** as the core component
  - Alternative to token self attention that **operates on the feature dimension**
  - **Eliminates** need for **expensive computation** of quadratic attention maps
- XCiT exhibits **strong image classification performance**
  - On par with state-of-the-art transformer models
  - **Robust to changing image resolutions**
- XCiT acts as a **very effective backbone** for dense prediction tasks such as **object detection**, **instance and semantic segmentation**
- XCiT is a strong backbone for **self-supervised learning**

Thank You