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, Hervé Jégou

NeurIPS 2021

Inria

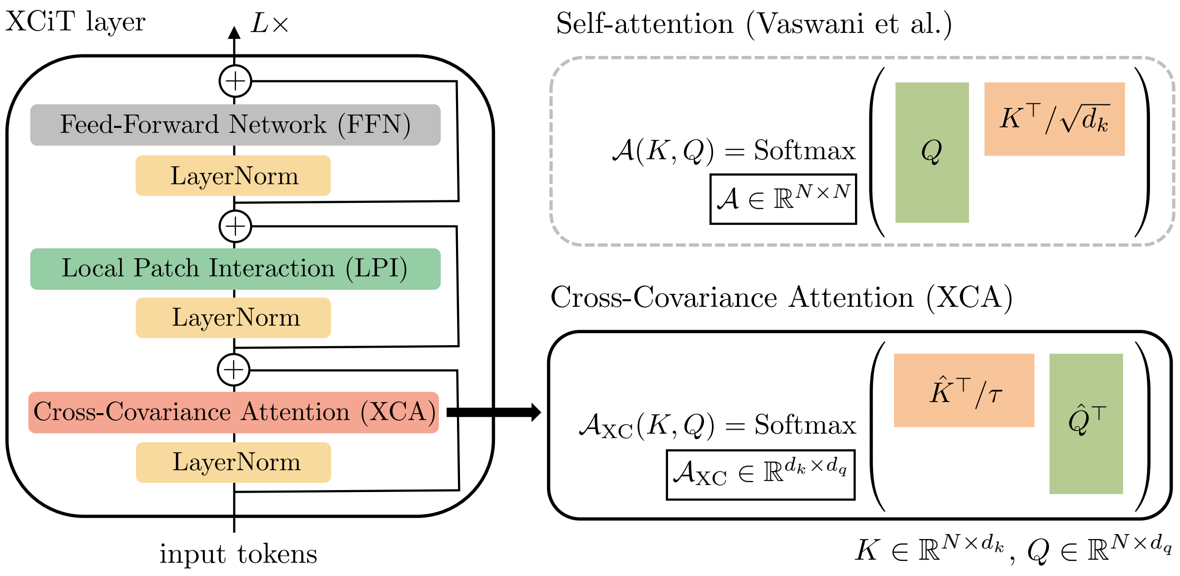
FACEBOOK AI

What is XCiT ?

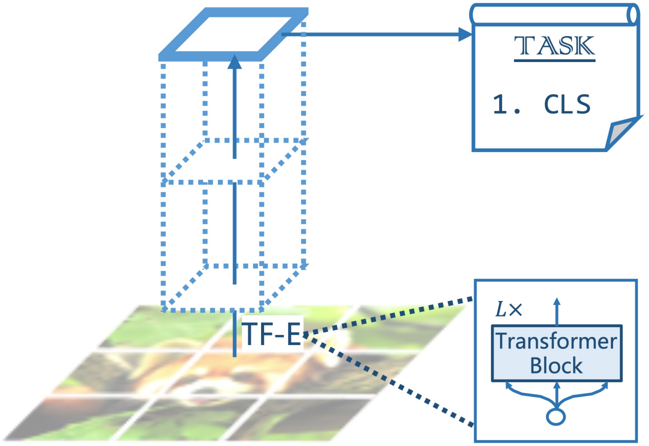
XCiT is a new form of Vision Transformers with Cross-Covariance Attention (XCA) as its core operation.

XCiT has linear complexity in image size (i.e. number of patches). It achieves a balance between the strong performance of ViT models and the flexiblity and scalibility of ConvNets in dealing of variable sized images.

Due to the favorable properties of XCiT, it exhibits strong performances for a variety of computer vision tasks, including dense prediction task like detection and segmentation.



Background: Vision Transformers



Vision Transformers (ViT) have shown a very strong performance for image classification using self-attention as the core operation in a convolutional-free model (aside from the linear projection).

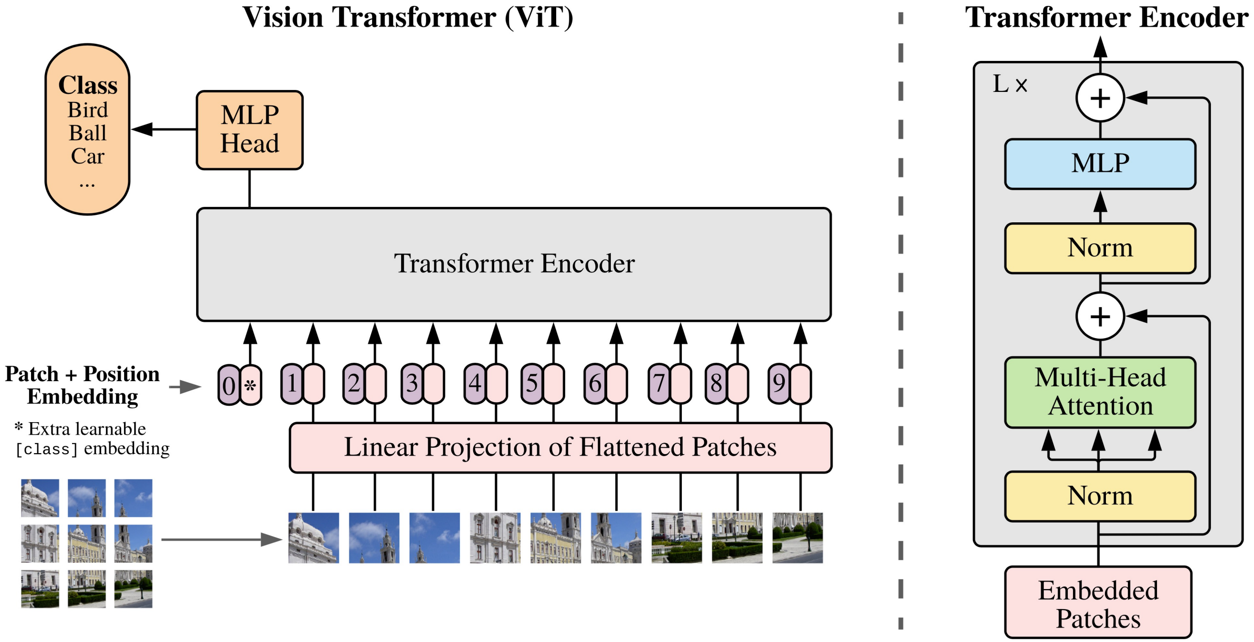
Self-Attention

N^2

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Using 16x16 patches

- ImageNet 224 images: N=196
- COCO 1300x800 image: N=4100

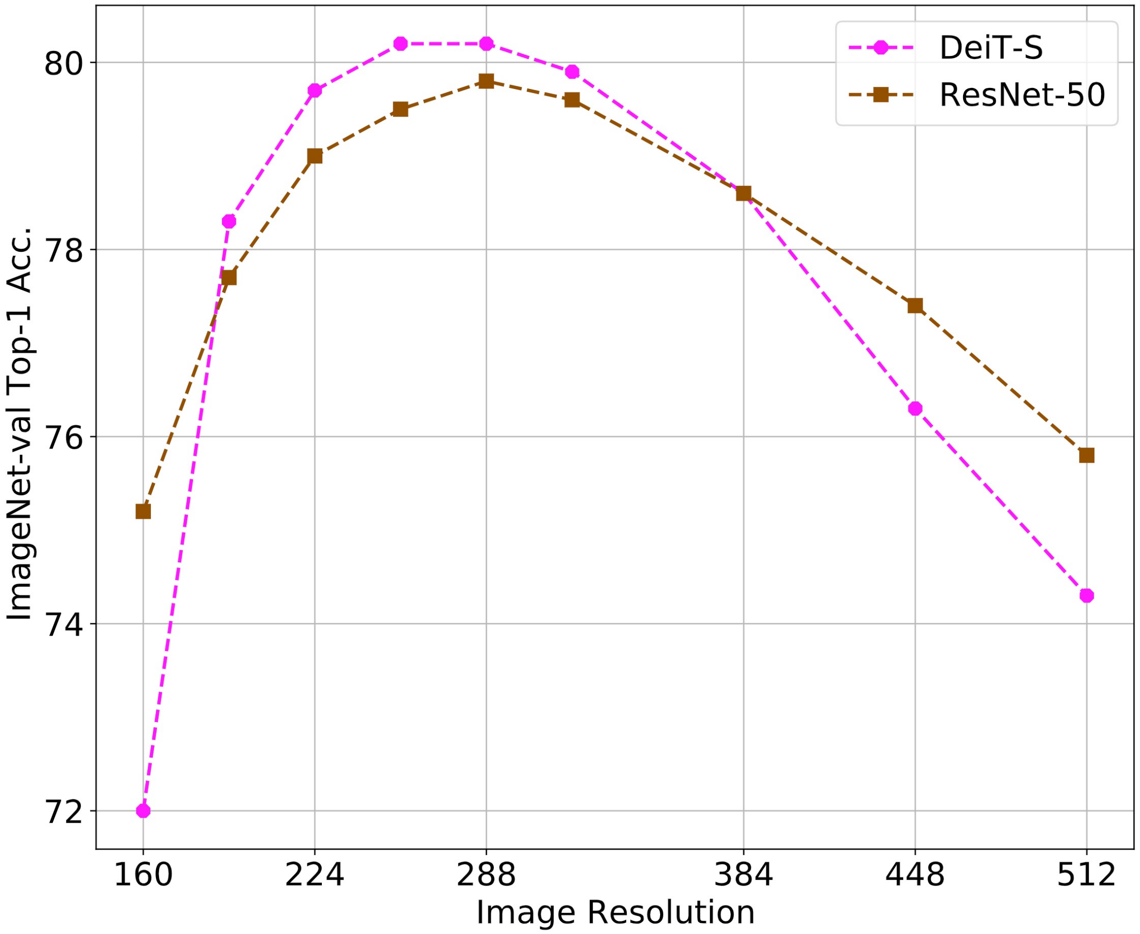


Background: Vision Transformers

ViT-Small (DeiT) achieves a higher performance compared to ResNet-50 on a standard ImageNet benchmark using 224 images.

However, we can notice that when ViT is tested using a different resolution, it quickly drops in performance as we move away from the train resolution. This can be harmful for tasks requiering processing of variable resolution images (e.g. Object Detection)

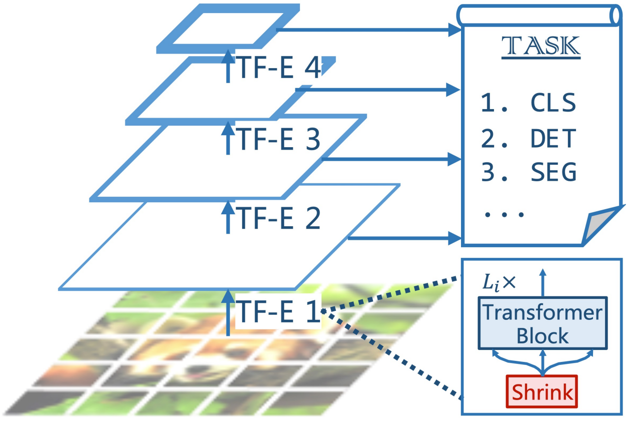
On the other hand ResNet-50 shows a better robsutness to changes in resolution.



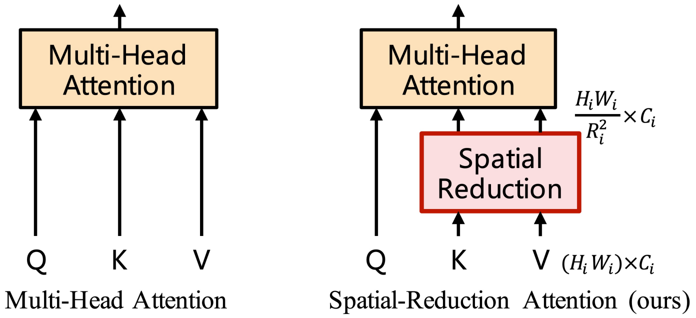
Concurent Work: Efficient Transformers

Pyramid Vision Transformer

a) Multi-Scale

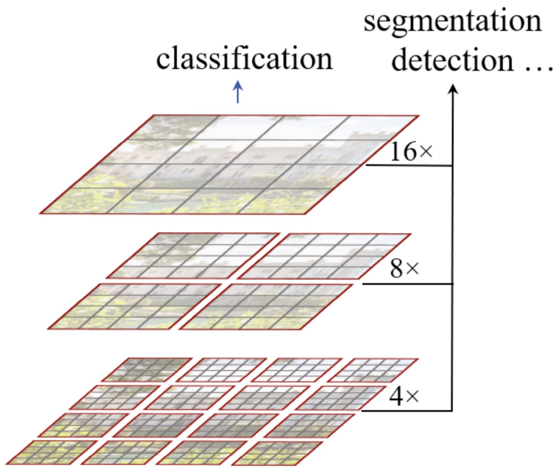


b) Approximate Attention

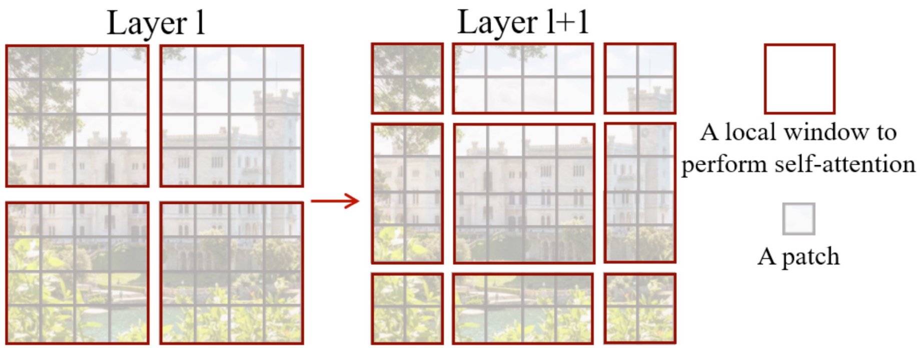


Swin Transformer

a) Multi-Scale



b) Approximate Attention



Motivation: Cross-Covariance Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\boxed{QK^T}}{\sqrt{d_k}}\right)V$$

The inner product between the Queries and Keys resmebles the Gram Matrix *G.*
In the special case where the projection matrices are identity, this relationship is exact.

$$QK^\top = XW_qW_k^\top X^\top \qquad G = XX^\top$$

The Gram and the Covariance matrices have a strong relationship that have been used for efficient computation of Principle components (PCA).

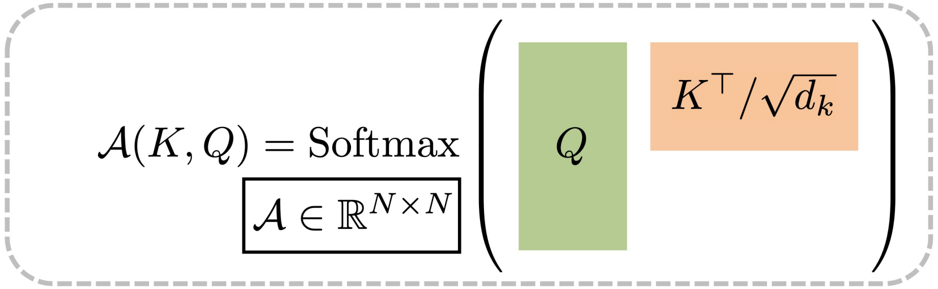
$$G = XX^\top \qquad C = X^\top X$$

The non-zero part of the eigenspectrum of the Gram and covariance matrix are equivalent, and the eigenvectors of C and G can be computed in terms of each other.

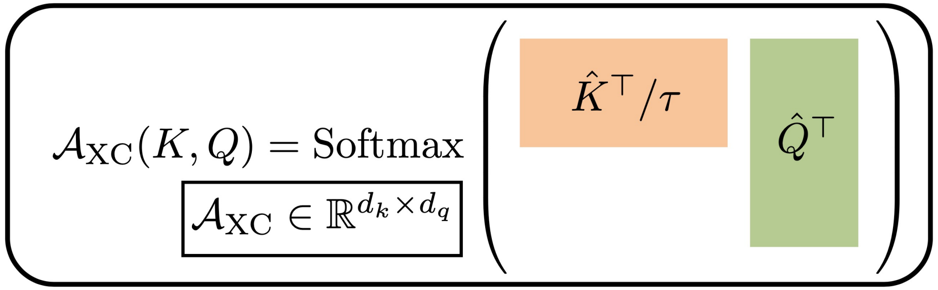
If V is the eigenvectors of G , then U the eigenvectors of C:

$$U = XV$$

Self-attention (Vaswani et al.)



Cross-Covariance Attention (XCA)



$K \in \mathbb{R}^{N \times d_k}$, $Q \in \mathbb{R}^{N \times d_q}$

Motivation: Cross-Covariance Attention

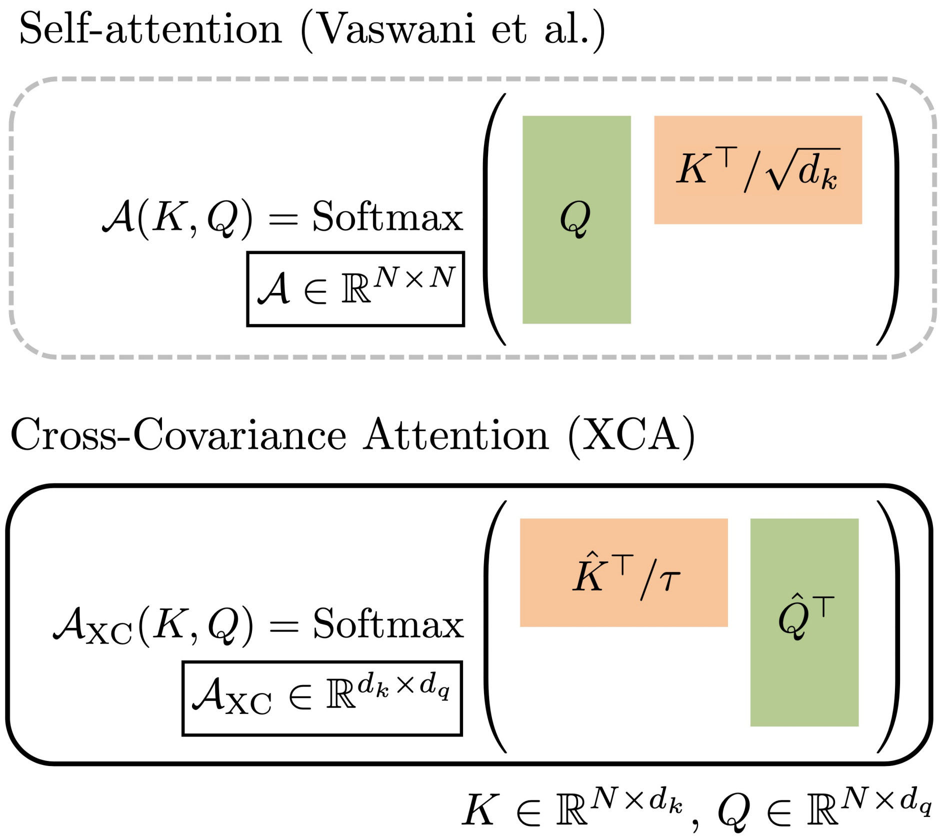
The covariance matrix has a complexity of d^2, we can study using attention over the covariance matrix as an alternative for the Gram based attention.

$$G = XX^\top \qquad C = X^\top X$$

$$QK^\top = XW_qW_k^\top X^\top \qquad K^\top Q = W_k^\top X^\top X W_q$$

Intuitively, we can think of cross-covariance attention as:

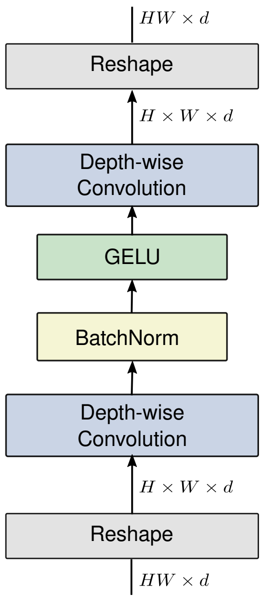
- Dynamically generating 1D filters based on the feature statistics across patches
- An advanced, attention-based version of Squeeze and Excitation

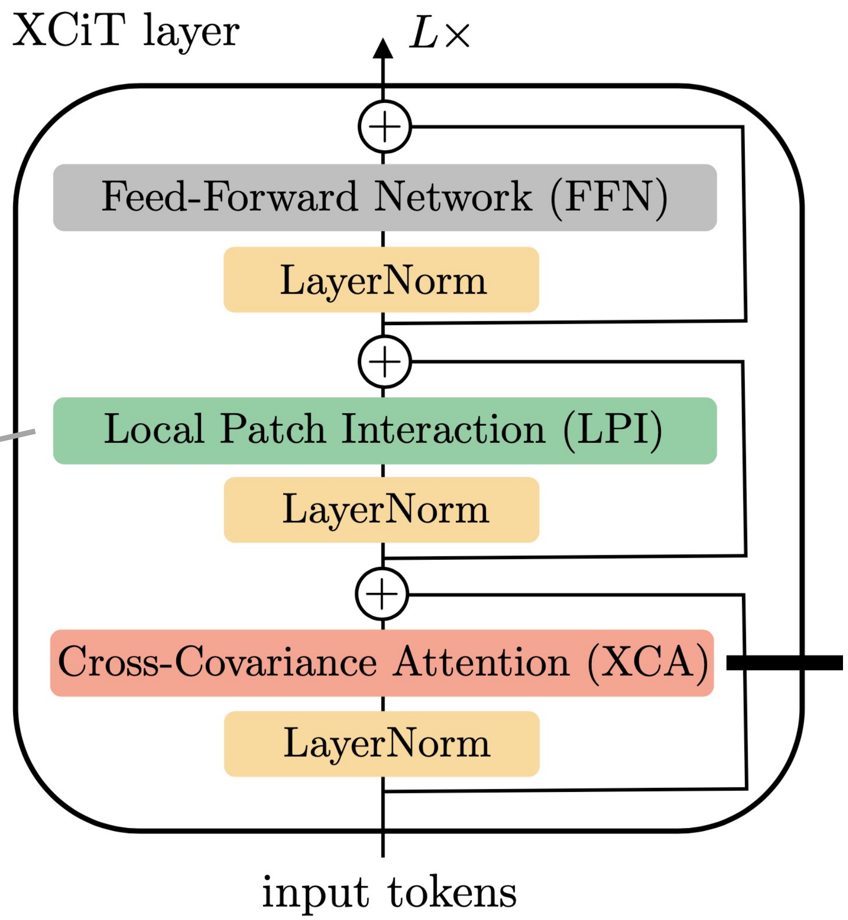


Cross-Covariance Image Transformer

We build the XCiT model with XCA at its core

- XCiT has a columnar structure with a consistent scale for the features from start to end.
- The linear projection of patches is replaced with a Convolutional based patch projection (similar to LeViT)
- We use the same FFN and LayerNorm setup as ViT.
- Since XCA only allows **Implicit** communication across patches. We add a Local Patch Interaction (LPI) module which consists of a light-weight depth-wise 3x3 Conv.





XCiT family of models

Model	Depth	d	#heads	#params	GFLOPs @224/16	GFLOPs @384/8	ImageNet-1k-val top-1 acc. (%) @224/16	ImageNet-1k-val top-1 acc. (%) @224/16Υ	ImageNet-1k-val top-1 acc. (%) @384/8Υ ↑
XCiT-N12	12	128	4	3M	0.5	6.4	69.9	72.2	77.8
XCiT-T12	12	192	4	7M	1.2	14.3	77.1	78.6	82.4
XCiT-T24	24	192	4	12M	2.3	27.3	79.4	80.4	83.7
XCiT-S12	12	384	8	26M	4.8	55.6	82.0	83.3	85.1
XCiT-S24	24	384	8	48M	9.1	106.0	82.6	83.9	85.6
XCiT-M24	24	512	8	84M	16.2	188.0	82.7	84.3	85.8
XCiT-L24	24	768	16	189M	36.1	417.9	82.9	84.9	86.0

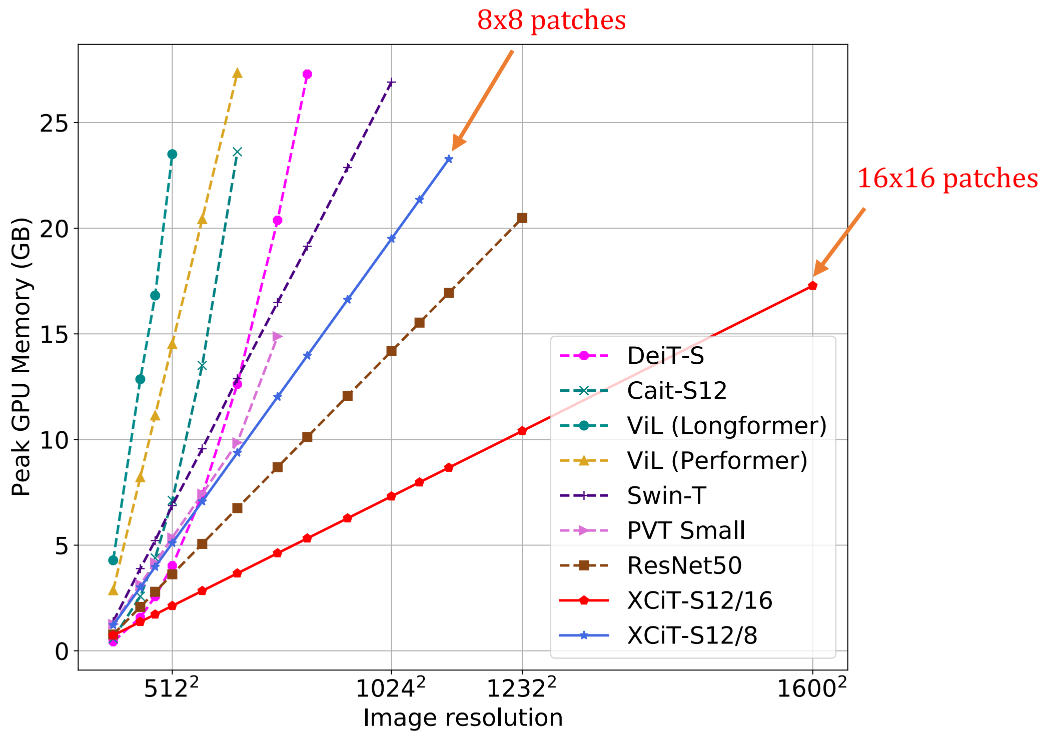
Based on the XCiT architecture, we designed a family of models with different trade-offs in accuracy, parameter count and FLOPS. The design parameters are:

- Number of Layers ∈ [12, 24]
- Dimensionality of the patch embeddings ∈ [128, 192, 384, 512, 768]
- Number of heads

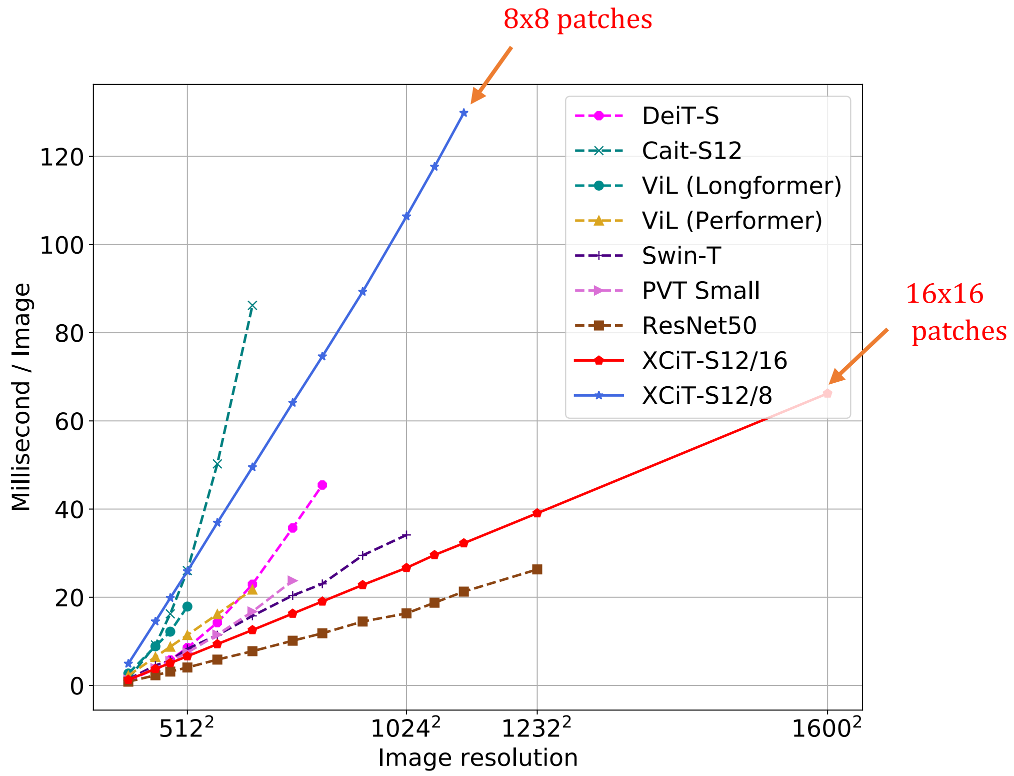
Since XCiT has linear complexity in number of patches, it allows for more fine-grained sampling of the patches. We experiment with 8x8 patches in addition to the 16x16 ones.

Using 8x8 patches and 384 image we can achieve a strong performance of 86.0% on IN-1k top-1, outperforming SoTA methods under the same number of parameters.

XCiT: Memory and Throughput



Memory usage (Inference)



Millisecond/Image (Inference)

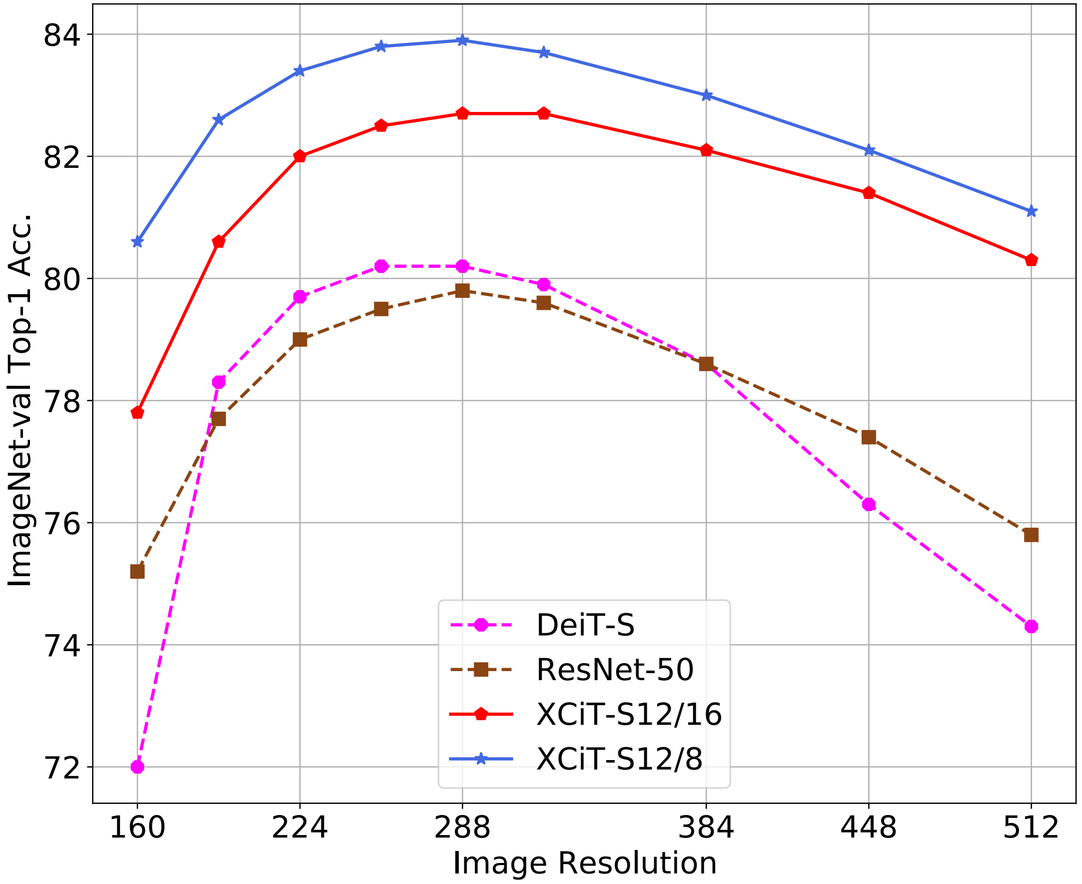
Model	#params	ImNet	Image Resolution							
	($\times 10^6$)	Top-1	224^2		384^2		512^2		1024^2	
		@224	im/sec	mem (MB)	im/sec	mem (MB)	im/sec	mem (MB)	im/sec	mem (MB)
ResNet-50	25	79.0	1171	772	434	2078	245	3618	61	14178
DeiT-S	22	79.9	974	433	263	1580	116	4020	N/A	OOM
CaiT-S12	26	80.8	671	577	108	2581	38	7117	N/A	OOM
PVT-Small	25	79.8	777	1266	256	3142	134	5354	N/A	OOM
Swin-T	29	81.3	704	1386	220	3890	120	6873	29	26915
XCiT-S12/16	26	82.0	781	731	266	1372	151	2128	37	7312

XCiT: Variable Sized Images

The cross-covariance attention, in particular the softmax operation, operates over a constant number of entities (i.e. d channels), regardless what is the image size.

On the other hand, Gram-based self-attention can suffer from a shift in statistics when the image size changes.

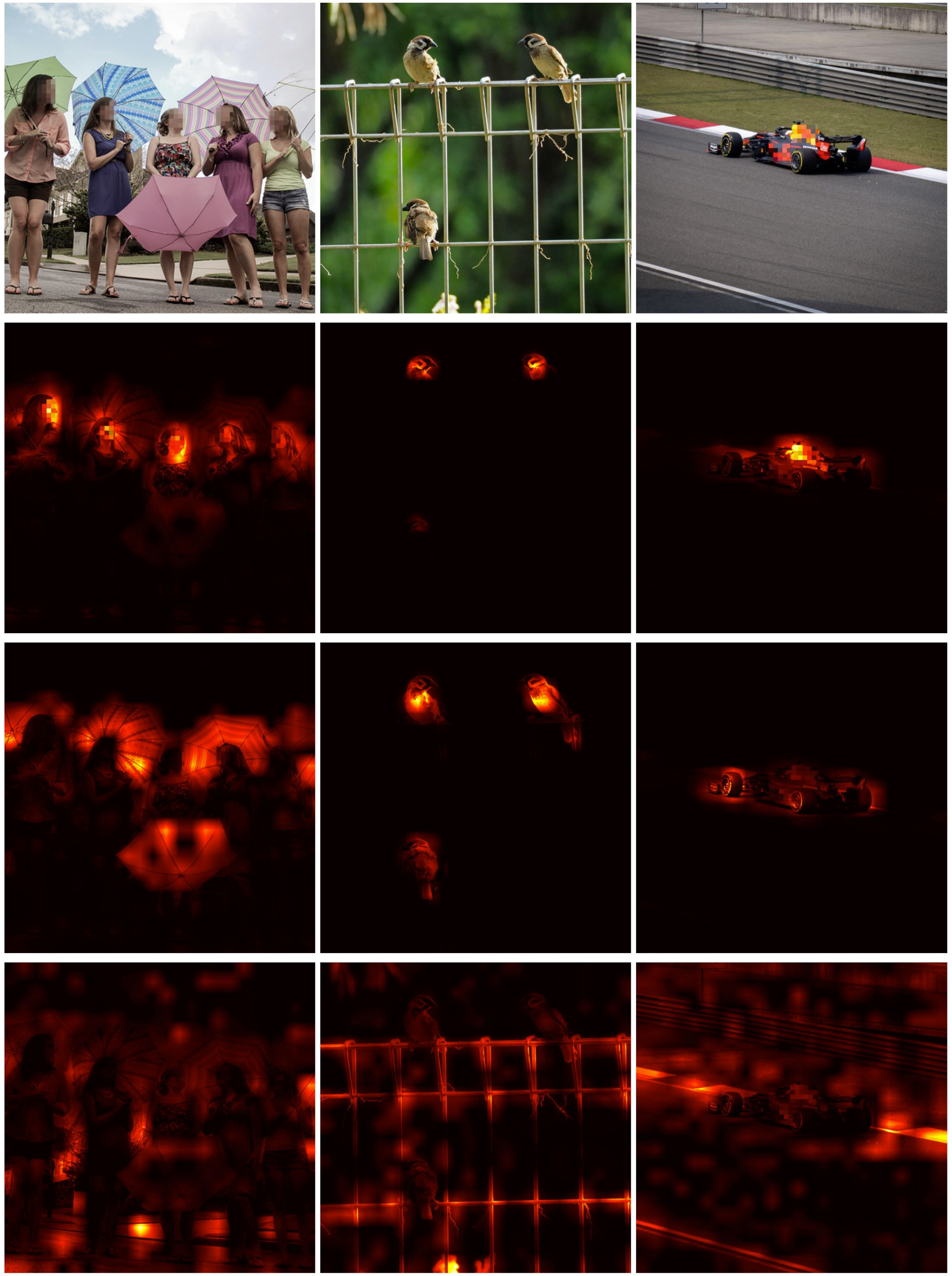
We can see that XCiT has a much better behaviour compared to ViT/DeiT w.r.t the drop in performance as the test image resolution changes. The behaviour matches or exceeds that of ConvNets (ResNet-50).



XCiT: Visualizations

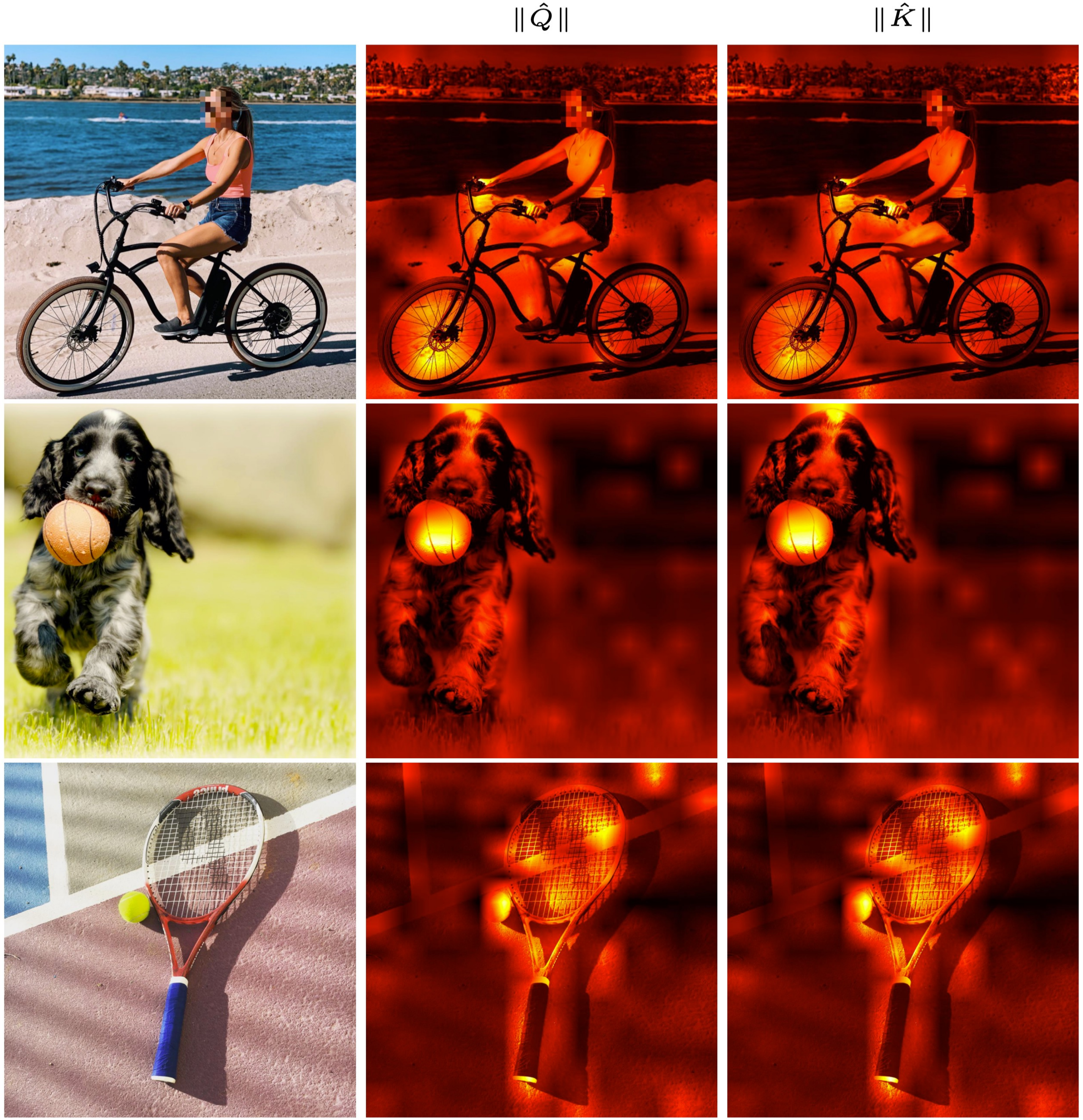
Visualization of the CLS attention layer (Gram-based)

- Every head (rows) attends to semantically coherent salient regions in the image
- Some patterns emerge, such that the head salient to humans heads, highlights birds heads as well. However, when such a pattern is not present, it can dedicate its capacity towards a different salient region like a car cockpit.



XCiT: Visualizations

We can also visualize the spatial regions contributing most to the cross-covariance matrix by simply computing the magnitude of each patch embedding in the Keys or the queries



Results: Image Classification

- XCiT outperforms/matches all other previous and concurrent methods when comparing models of similar parameter counts, including CaiT and NFNets.
- We can observe a strong boost in performance when the 8x8 patch size is used, which is only enabled by the linear complexity of XCiT.
- The gain in performance due to the 8x8 patches is accompanied by higher FLOPS.

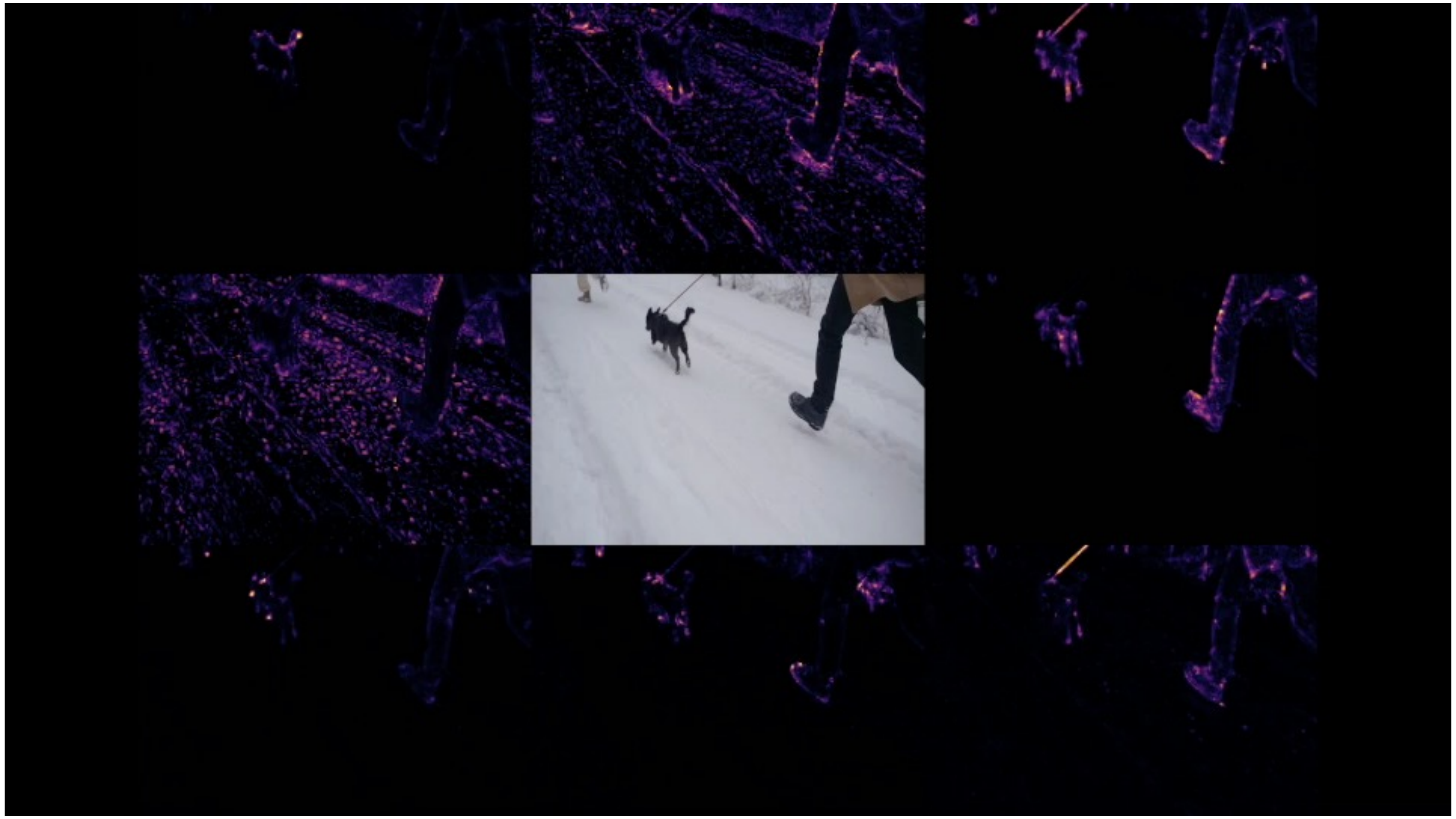
Model	#params	FLOPs	Res.	ImNet	V2
EfficientNet-B5 RA [18]	30M	9.9B	456	83.7	_
RegNetY-4GF [53]	21M	4.0B	224	80.0	72.4
DeiT-Sϒ [65]	22M	4.6B	224	81.2	68.5
Swin-T [44]	29M	4.5B	224	81.3	_
CaiT-XS24ϒ ↑ [68]	26M	19.3B	384	84.1	74.1
XCiT-S12/16ϒ	26M	4.8B	224	83.3	72.5
XCiT-S12/16ϒ ↑	26M	14.3B	384	84.7	74.1
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**

Results: SSL with DINO

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**	-

Results: SSL with DINO

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**	-



Results: Ablations

- We notice that the convolutional patch projection imporves the performance strongly for 16x16 patch models, but the impact is smaller for 8x8 patch models
- The LPI module improves the performance by 1.2%. On the other hand, the model without XCA has a weak performance of 75.9%
- We notice that we have very unstable training the L2-Normalization and often our training collapses.
- The Learned temperature parameter has a positive small improvement to the performance with no overhead.

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

Results: Object detection w/ COCO

- XCiT uses a columnar structure with only one scale for all layers.
- To obtain multiple scale features for FPN, we use:
  - Maxpooling to obtain lower resolution features.
  - Transposed Convolution to obtain the higher resolution feature maps.
- We show that having a pyramidal structure is not a <u>necessity</u> for adapting transformers for dense prediction tasks.
- All our models uses a Mask R-CNN framework with XCiT only replacing the trunk. Models are trained for the standard 3x schedule.
- XCiT outperforms PVT and ViL across all operating points. It provides a competitive performance with Swin, where XCiT provides a better performance for smaller capacity models and Swin marginally improving the performance for the larger sized model.

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**

Results: Semantic Segmentation w/ ADE20k

- Uses the same FPN components as object detection
- XCiT outperforms ResNets, PVT, ViL and Swin for all operating points and using two different decoders.

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**

Object detection and instance segmentation of Ultra-High Res images (6000x4000)

Summary

- XCiT is a new vision transformer with linear complexity in image size, providing a large saving in terms of memory compared to recent vision transformers.
- XCiT achieves a balance between the strong performance of transformers and the flexibility of ConvNets.
- XCiT exhibits a strong performance on a variety of computer vision tasks including SSL, detection and segmentation.
- Code and weights available: https://github.com/facebookresearch/xcit