Visual Attention Network

Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng and Shi-Min Hu

Motivation

TABLE 1
Desirable properties belonging to convolution, self-attention and LKA.

Properties	Convolution	Self-Attention	LKA
Local Receptive Field	✓	✗	✓
Long-range Dependence	✗	✓	✓
Spatial Adaptability	✗	✓	✓
Channel Adaptability	✗	✗	✓
Computational complexity	$\mathcal{O}(n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$

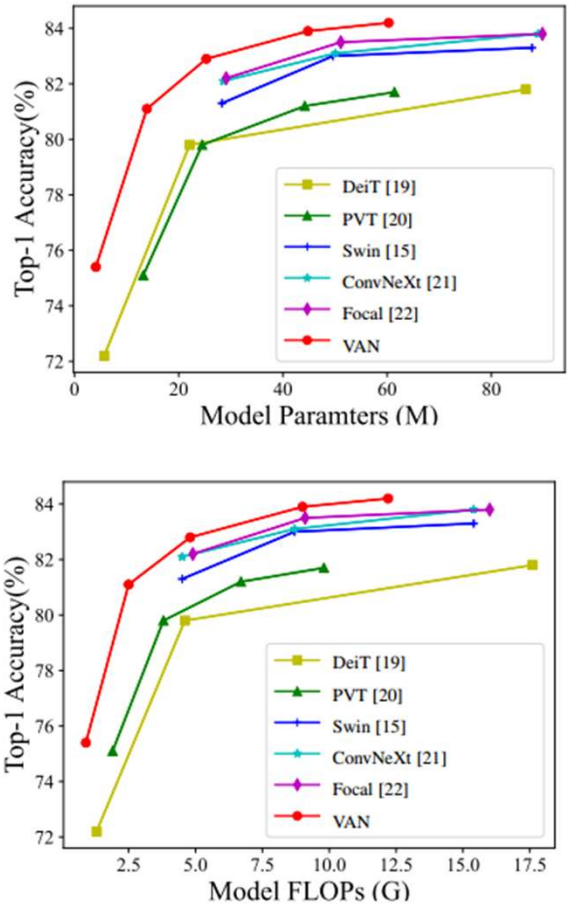
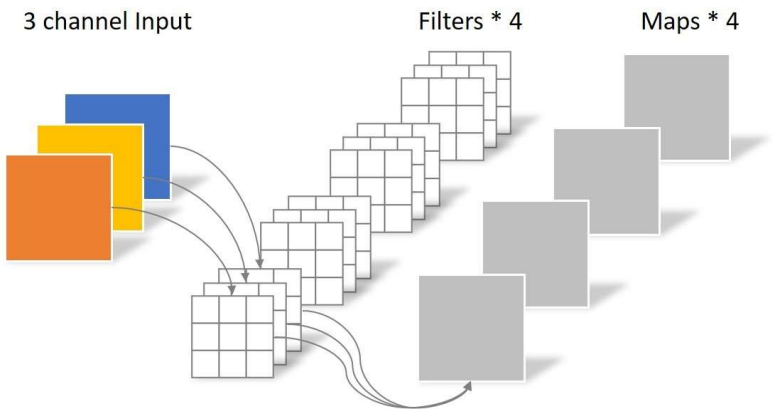


Fig. 1. Results of different models on ImageNet-1K validation set. Comparing the performance of recent models DeiT [19], PVT [20], Swin Transformer [15], ConvNeXt [21], Focal Transformer [22] and our VAN. Above: Accuracy-Parameters trade-off diagram. Under: Accuracy-FLOPs trade-off diagram.

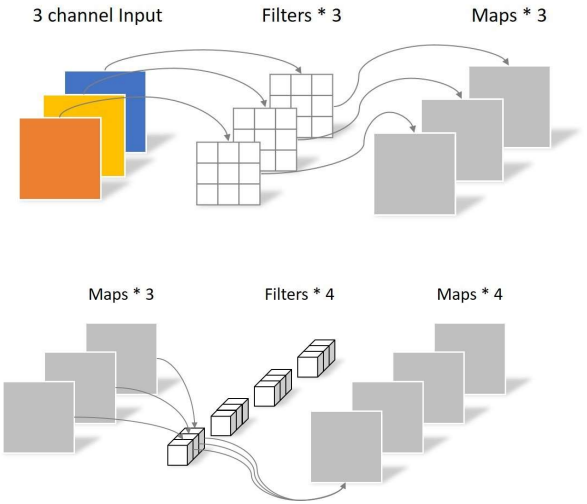
depthwise separable convolution



标准卷积

卷积核的尺寸是 $k \times k \times M$ 一共有 N 个

所以标准卷积的参数量是 $\text{Params: } K \times K \times M \times N$



深度可分离卷积

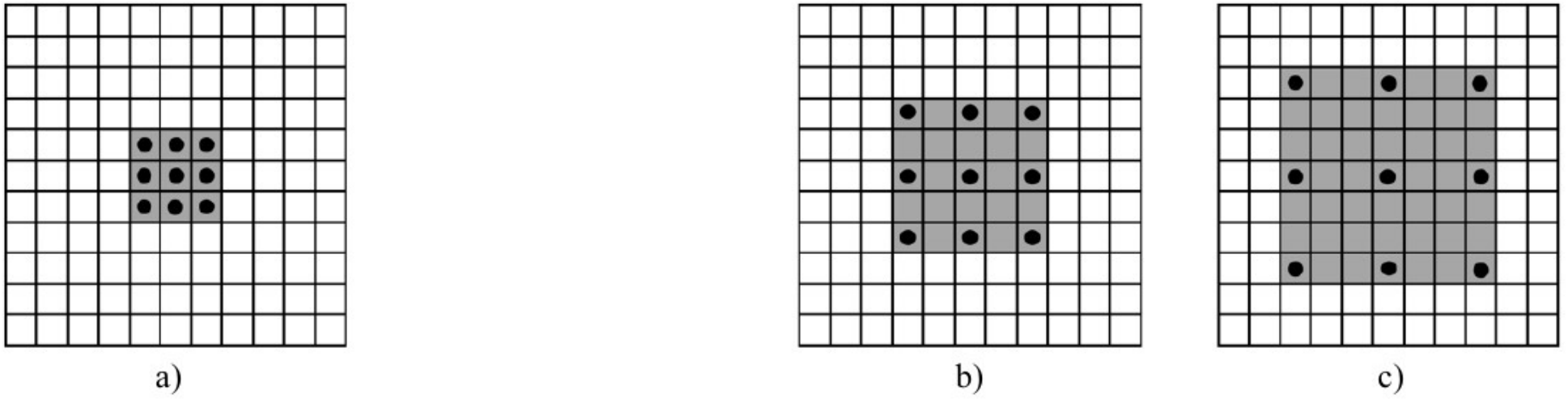
$\text{depthwise - convolution}$ 其参数量为: $K \times K \times M$

$pointwise - convolution$ 参数量是: $M \times N$

总参数量是: $K \times K \times M + M \times N$

$$\frac{K \times K \times M + M \times N}{K \times K \times M \times N} = \frac{1}{N} + \frac{1}{K^2}$$

Dilated/Atrous Convolution



large kernel attention, LKA

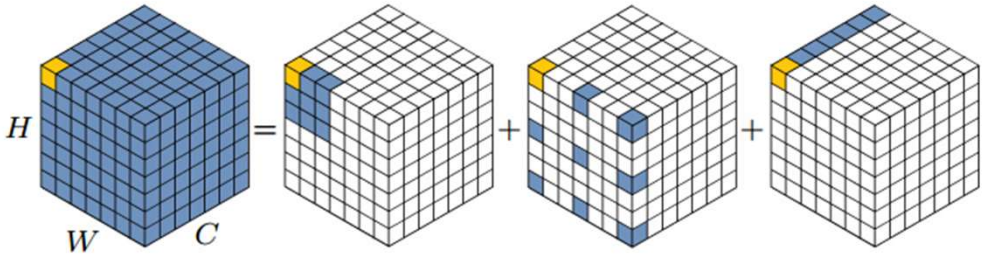


Fig. 2. Decomposition diagram of large-kernel convolution. A standard convolution can be decomposed into three parts: a depth-wise convolution (DW-Conv), a depth-wise dilation convolution (DW-D-Conv), and a pointwise convolution (1×1 Conv). The colored grids represent the location of convolution kernel and the yellow grid means the center point. The diagram shows that a 13×13 convolution is decomposed into a 5×5 depth-wise convolution, a 5×5 depth-wise dilation convolution with dilation rate 3, and a pointwise convolution. Note: zero paddings are omitted in the above figure.

Specifically. we can decompose a $K \times K$ convolution into a $\lceil \frac{K}{d} \rceil \times \lceil \frac{K}{d} \rceil$ depth-wise dilation convolution with dilation d, a $(2d-1) \times (2d-1)$ depth-wise convolution and a 1×1 convolution.

$$Attention = \text{Conv}_{1\times1}(\text{DW-D-Conv}(\text{DW-Conv}(F))), \quad (1)$$

$$Output = Attention \otimes F. \quad (2)$$

Here, $F \in \mathbb{R}^{C\times H\times W}$ is the input feature. $Attention \in \mathbb{R}^{C\times H\times W}$ denotes attention map. The value in attention map indicates the importance of each feature. $\otimes$ means element-wise product. Different from common attention methods, LKA dose not require an additional normalization function like sigmoid and softmax, which is demonstrated in Tab. 3. We also believe the

Visual Attention Network

TABLE 5
The detailed setting for different versions of the VAN. e.r. represents expansion ratio in the feed-forward network.

stage	output size	e.r.	VAN-						
			B0	B1	B2	B3	B4	B5	B6
1	$\frac{H}{4} \times \frac{W}{4} \times C$	8	$C = 32$ $L = 3$	$C = 64$ $L = 2$	$C = 64$ $L = 3$	$C = 64$ $L = 3$	$C = 64$ $L = 3$	$C = 96$ $L = 3$	$C = 96$ $L = 6$
2	$\frac{H}{8} \times \frac{W}{8} \times C$	8	$C = 64$ $L = 3$	$C = 128$ $L = 2$	$C = 128$ $L = 3$	$C = 128$ $L = 5$	$C = 128$ $L = 6$	$C = 192$ $L = 3$	$C = 192$ $L = 6$
3	$\frac{H}{16} \times \frac{W}{16} \times C$	4	$C = 160$ $L = 5$	$C = 320$ $L = 4$	$C = 320$ $L = 12$	$C = 320$ $L = 27$	$C = 320$ $L = 40$	$C = 480$ $L = 24$	$C = 384$ $L = 90$
4	$\frac{H}{32} \times \frac{W}{32} \times C$	4	$C = 256$ $L = 2$	$C = 512$ $L = 2$	$C = 512$ $L = 3$	$C = 512$ $L = 3$	$C = 512$ $L = 3$	$C = 768$ $L = 3$	$C = 768$ $L = 6$
Parameters (M)			4.1	13.9	26.6	44.8	60.3	90.0	200
FLOPs (G)			0.9	2.5	5.0	9.0	12.2	17.2	38.4

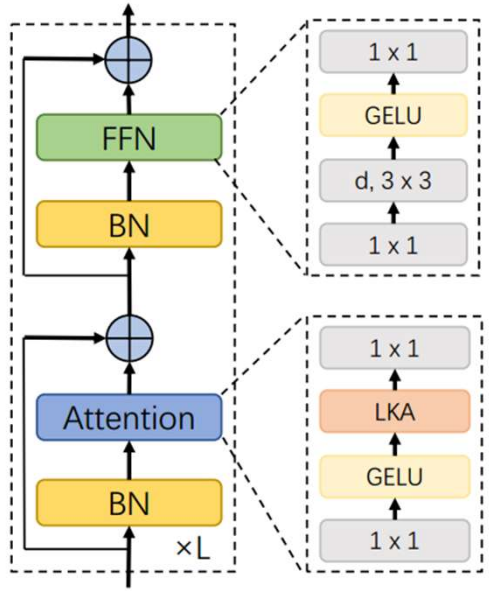


Fig. 4. A stage of VAN. d means depth wise convolution. k × k denotes k × k convolution.

Visual Attention Network

TABLE 2
Number of parameters for different forms of a 21×21 convolution. For instance, when the number of channels $C = 32$, standard convolution and MobileNet decomposition use $133\times$ and $4.5\times$ more parameters than our decomposition respectively.

	Standard Convolution	Decomposition Type	
		MobileNet [6]	Ours
C=32	451,584	15,136	3,392
C=64	1,806,336	32,320	8,832
C=128	7,225,344	72,832	25,856
C=256	28,901,376	178,432	84,480
C=512	115,605,504	487,936	300,032

$$P(K,d) = C(\lceil \frac{K}{d} \rceil^2 \times C + (2d-1)^2) + C^2, \quad (3)$$

$$F(K,d) = P(K,d) \times H \times W. \quad (4)$$

Ablation Studies

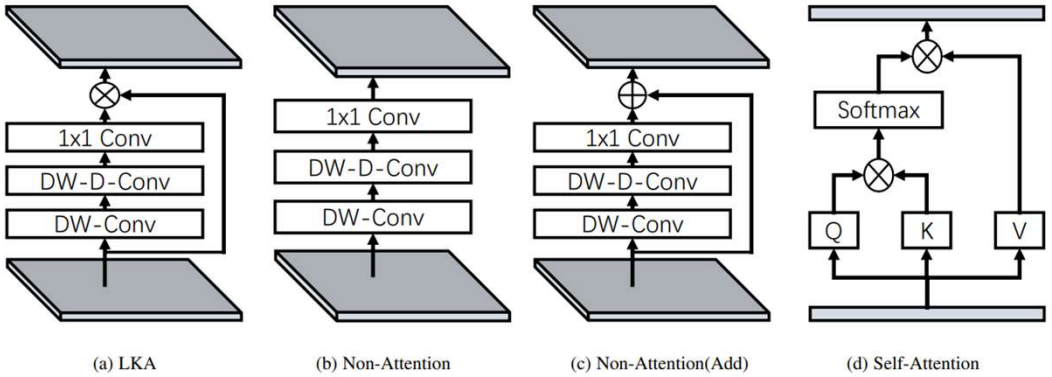


Fig. 3. The structure of different modules: (a) the proposed Large Kernel Attention (LKA); (b) non-attention module; (c) replace multiplication in LKA with addition ; (d) self-attention. It is worth noting that (d) is designed for 1D sequences.

TABLE 3
Ablation study of different modules in LKA. Top-1 accuracy (Acc) on ImageNet validation set suggest that each part is critical. w/o Attention means we adopt Fig. 3(b).

VAN-B0	Params. (M)	FLOPs(G)	Acc(%)
w/o DW-Conv	4.1	0.9	74.9
w/o DW-D-Conv	4.0	0.9	74.1
w/o Attention	4.1	0.9	74.3
w/o Attention (Add)	4.1	0.9	74.6
w/o 1 × 1 Conv	3.8	0.8	74.6
w/ Sigmoid	4.1	0.9	75.2
VAN-B0	4.1	0.9	75.4

- DW-Conv -- 性能下降0.5% (74.9%对75.4%)
- DW-D-Conv -- 分类性能下降1.3% (74.1%对75.4%)
- 注意机制 --约1.1% (74.3% vs. 75.4%)的改进。
- 1 × 1 Conv --提高了0.8% (74.6% vs. 75.4%)
- Sigmoid函数 --去除 Sigmoid 实现了0.2% (75.4% vs. 75.2%)的改进

Ablation Studies

TABLE 6
Ablation study of different kernel size K in LKA. Acc(%) means Top-1 accuracy on ImageNet validation set.

Method	K	Dilation	Params. (M)	GFLOPs	Acc(%)
VAN-B0	7	2	4.03	0.85	74.8
VAN-B0	14	3	4.07	0.87	75.3
VAN-B0	21	3	4.11	0.88	75.4
VAN-B0	28	4	4.14	0.90	75.4

Comparison with Existing Methods

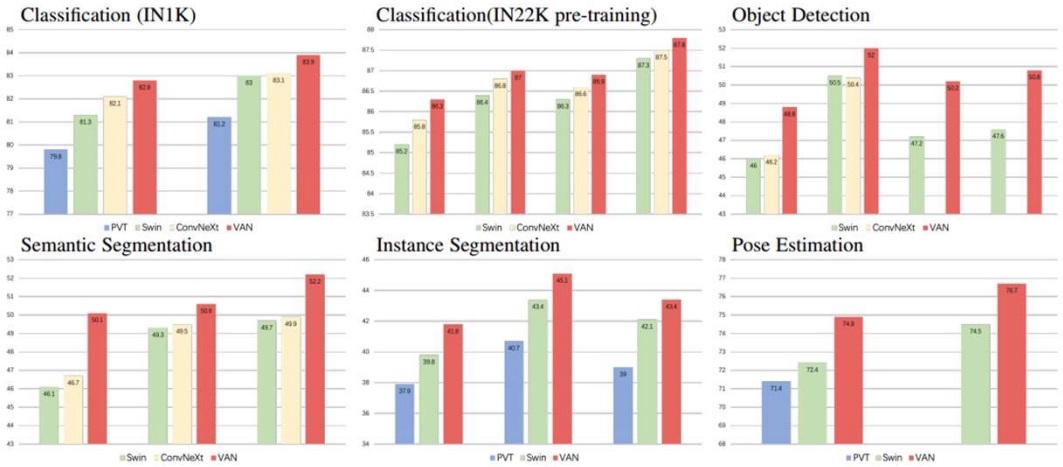


Fig. 6. Comparing with similar level PVT [20], Swin Transformer [15] and ConvNeXt [21] on various tasks, including image classification, object detection, semantic segmentation, instance segmentation and pose estimation.

ConvNeXt-T提高了0.7%(82.8%对82.1%)

VAN- b2比Swin-T 高1.5% (82.8% vs. 81.3%)。

VAN-B2比 gMLP-S 高出3.2%(82.8%比79.6%)

TABLE 7
Compare with the state-of-the-art methods on ImageNet validation set. Params means parameter. GFLOPs denotes floating point operations. Top-1 Acc represents Top-1 accuracy.FLOPs is

Method	Params. (M)	GFLOPs	Top-1 Acc (%)
PVTv2-B0 [80]	3.4	0.6	70.5
T2T-ViT-7 [54]	4.3	1.1	71.7
DeiT-Tiny/16 [19]	5.7	1.3	72.2
TNT-Ti [97]	6.1	1.4	73.9
VAN-B0	4.1	0.9	**75.4**
ResNet18 [5]	11.7	1.8	69.8
PVT-Tiny [20]	13.2	1.9	75.1
PoolFormer-S12 [98]	11.9	2.0	77.2
PVTv2-B1 [80]	13.1	2.1	78.7
VAN-B1	13.9	2.5	**81.1**
ResNet50 [5]	25.6	4.1	76.5
ResNeXt50-32x4d [7]	25.0	4.3	77.6
RegNetY-4G [99]	21.0	4.0	80.0
DeiT-Small/16 [19]	22.1	4.6	79.8
T2T-ViT$_t$-14 [54]	21.5	6.1	81.7
PVT-Small [20]	24.5	3.8	79.8
TNT-S [97]	23.8	5.2	81.3
ResMLP-24 [71]	30.0	6.0	79.4
gMLP-S [72]	20.0	4.5	79.6
Swin-T [15]	28.3	4.5	81.3
PoolFormer-S24 [98]	21.4	3.6	80.3
Twins-SVT-S [100]	24.0	2.8	81.7
PVTv2-B2 [80]	25.4	4.0	82.0
Focal-T [22]	29.1	4.9	82.2
ConvNeXt-T [21]	28.6	4.5	82.1
VAN-B2	26.6	5.0	**82.8**
ResNet101 [5]	44.7	7.9	77.4
ResNeXt101-32x4d [7]	44.2	8.0	78.8
Mixer-B/16 [69]	59.0	11.6	76.4
T2T-ViT$_t$-19 [54]	39.2	9.8	82.4
PVT-Medium [20]	44.2	6.7	81.2
Swin-S [15]	49.6	8.7	83.0
ConvNeXt-S [15]	50.1	8.7	83.1
PVTv2-B3 [80]	45.2	6.9	83.2
Focal-S [22]	51.1	9.1	83.5
VAN-B3	44.8	9.0	**83.9**
ResNet152 [5]	60.2	11.6	78.3
T2T-ViT$_t$-24 [54]	64.0	15.0	82.3
PVT-Large [20]	61.4	9.8	81.7
TNT-B [97]	66.0	14.1	82.8
PVTv2-B4 [80]	62.6	10.1	83.6
VAN-B4	60.3	12.2	**84.2**

Experiments

TABLE 4
Throughput of Swin transformer and VAN on RTX 3090.

Method	FLOPs(G)	Throughput (Imgs/s)	Acc(%)
Swin-T	4.5	821	81.3
Swin-S	8.7	500	83.0
Swin-B	15.4	376	83.5
VAN-B0	0.9	2140	75.4
VAN-B1	2.5	1420	81.1
VAN-B2	5.0	762	82.8
VAN-B3	9.0	452	83.9
VAN-B4	12.2	341	84.2

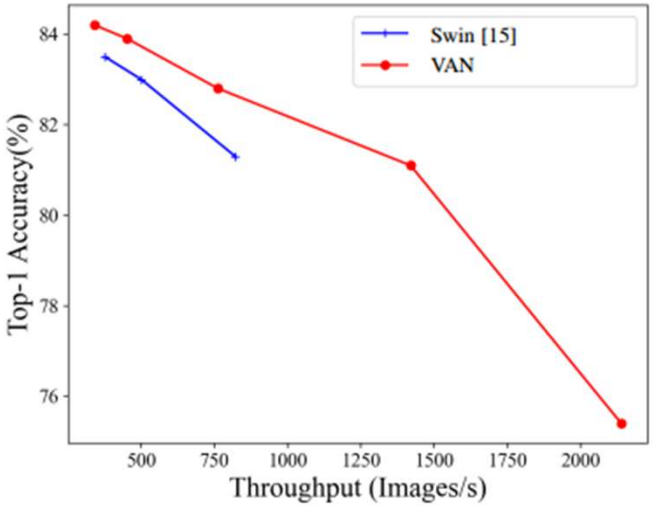


Fig. 5. Accuracy-Throughput Diagram. It claerly shows that VAN achieves a better trade-off than swin transformer [15].

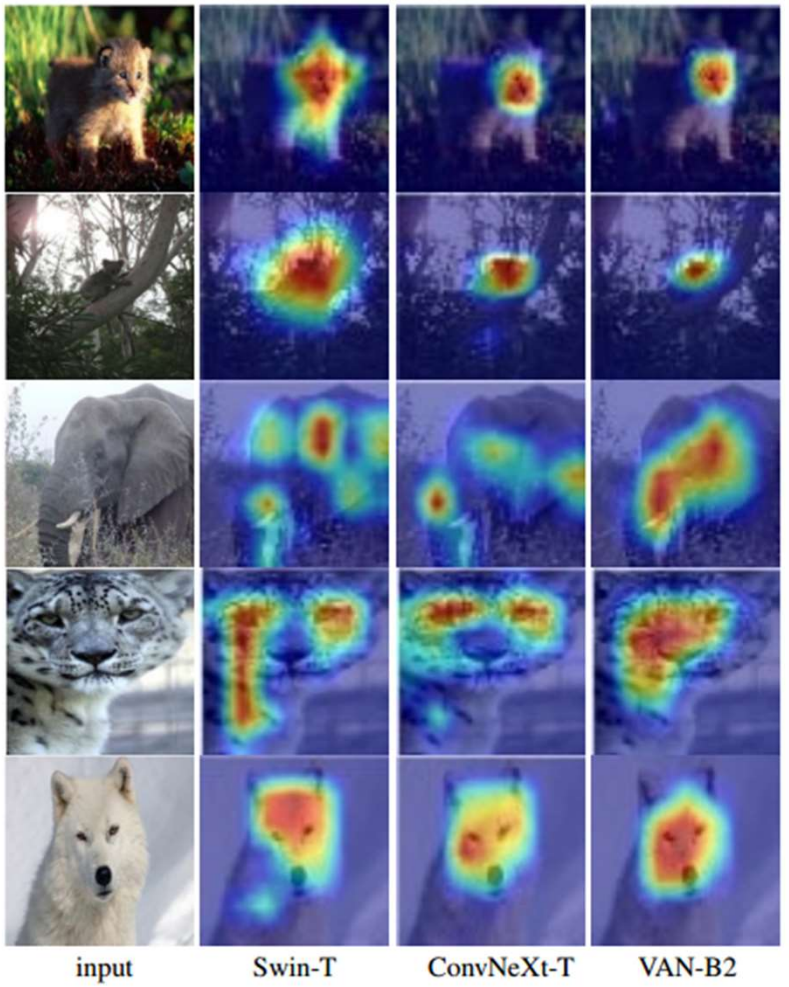


Fig. 7. Visualization results. All images come from different categories in ImageNet validation set. CAM is produced by using Grad-CAM [85]. We compare different CAMs produced by Swin-T [15], ConvNeXt-T [21] and VAN-B2.

Pre-training

TABLE 8
Compare with the state-of-the-art methods on ImageNet validation set. Params means parameter. GFLOPs denotes floating point operations. Top-1 Acc represents Top-1 accuracy.All models are pretrained on ImageNet-22K dataset.

Method	Params. (M)	Input size	GFLOPs	Top-1 Acc (%)
Swin-S [15]	50	224^2	8.7	83.2
ConvNeXt-S [21]	50	224^2	8.7	84.6
VAN-B4	60	224^2	12.2	**85.7**
ConvNeXt-S [21]	50	384^2	25.5	85.8
VAN-B4	60	384^2	35.9	**86.6**
Swin-B [15]	88	224^2	15.4	85.2
ConvNeXt-B [21]	89	224^2	15.4	85.8
VAN-B5	90	224^2	17.2	**86.3**
EffNetV2-L [101]	120	480^2	53.0	86.8
ViT-B/16 [13]	87	384^2	55.5	85.4
Swin-B [15]	88	384^2	47.0	86.4
ConvNeXt-B [21]	89	384^2	45.1	86.8
VAN-B5	90	384^2	50.6	**87.0**
Swin-L [15]	197	224^2	34.5	86.3
ConvNeXt-L [21]	198	224^2	34.4	86.6
VAN-B6	200	224^2	38.9	**86.9**
EffNetV2-XL [101]	208	480^2	94.0	87.3
CoAtNet-3 [102]	168	384^2	107.4	87.6
Swin-L [15]	197	384^2	103.9	87.3
ConvNeXt-L [21]	198	384^2	101.0	87.5
VAN-B6	200	384^2	114.3	**87.8**

Object Detection

TABLE 9
Object detection on COCO 2017 dataset. #P means parameter. RetinaNet 1× denotes models are based on RetinaNet [103] and we train them for 12 epochs.

Backbone	RetinaNet 1×						
	#P (M)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
VAN-B0	13.4	**38.8**	**58.8**	**41.3**	**23.4**	**42.8**	**50.9**
ResNet18 [5]	21.3	31.8	49.6	33.6	16.3	34.3	43.2
PoolFormer-S12 [20]	21.7	36.2	56.2	38.2	20.8	39.1	48.0
PVT-Tiny [20]	23.0	36.7	56.9	38.9	22.6	38.8	50.0
VAN-B1	23.6	**42.3**	**63.1**	**45.1**	**26.1**	**46.2**	**54.1**
ResNet50 [5]	37.7	36.3	55.3	38.6	19.3	40.0	48.8
PVT-Small [20]	34.2	40.4	61.3	43.0	25.0	42.9	55.7
PoolFormer-S24 [98]	31.1	38.9	59.7	41.3	23.3	42.1	51.8
PoolFormer-S36 [98]	40.6	39.5	60.5	41.8	22.5	42.9	52.4
VAN-B2	36.3	**44.9**	**65.7**	**48.4**	**27.4**	**49.2**	**58.7**
ResNet101 [5]	56.7	38.5	57.8	41.2	21.4	42.6	51.1
PVT-Medium [20]	53.9	41.9	63.1	44.3	25.0	44.9	57.6
VAN-B3	54.5	**47.5**	**68.4**	**51.2**	**30.9**	**52.1**	**62.4**

TABLE 10
Object detection and instance segmentation on COCO 2017 dataset. #P means parameter. Mask R-CNN 1× denotes models are based on Mask R-CNN [104] and we train them for 12 epochs. AP^b and AP^m refer to bounding box AP and mask AP respectively.

Backbone	Mask R-CNN 1×						
	#P (M)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
VAN-B0	23.9	**40.2**	**62.6**	**44.4**	**37.6**	**59.6**	**40.4**
ResNet18 [5]	31.2	34.0	54.0	36.7	31.2	51.0	32.7
PoolFormer-S12 [98]	31.6	37.3	59.0	40.1	34.6	55.8	36.9
PVT-Tiny [20]	32.9	36.7	59.2	39.3	35.1	56.7	37.3
VAN-B1	33.5	**42.6**	**64.2**	**46.7**	**38.9**	**61.2**	**41.7**
ResNet50 [5]	44.2	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Small [20]	44.1	40.4	62.9	43.8	37.8	60.1	40.3
PoolFormer-S24 [98]	41.0	40.1	62.2	43.4	37.0	59.1	39.6
PoolFormer-S36 [98]	50.5	41.0	63.1	44.8	37.7	60.1	40.0
VAN-B2	46.2	**46.4**	**67.8**	**51.0**	**41.8**	**65.2**	**44.9**
ResNet101 [5]	63.2	40.4	61.1	44.2	36.4	57.7	38.8
ResNeXt101-32x4d [7]	62.8	41.9	62.5	45.9	37.5	59.4	40.2
PVT-Medium [20]	63.9	42.0	64.4	45.6	39.0	61.6	42.1
VAN-B3	64.4	**48.3**	**69.6**	**53.3**	**43.4**	**67.0**	**46.8**

Semantic Segmentation

TABLE 13
Compare with the state-of-the-art methods on ADE20K validation set. Params means parameter. GFLOPs denotes floating point operations.All models are pretrained on ImageNet-22K dataset. We calculate FLOPs with input size 2560 × 640 for 640 input image and 2048 × 512 for 512 input image.

Method	Params. (M)	Input size	GFLOPs	mIoU
Swin-B [15]	121	640^2	1841	51.7
ConvNeXt-B [21]	122	640^2	1828	53.1
VAN-B5	117	512^2	1208	**53.9**
Swin-L [15]	234	640^2	2468	53.5
ConvNeXt-L [21]	235	640^2	2458	53.7
VAN-B6	231	512^2	1658	**54.7**

TABLE 12
Results of semantic segmentation on ADE20K [83] validation set. The upper and lower part are obtained under two different training/validation schemes following [98] and [15]. We calculate FLOPs with input size 512 × 512 for Semantic FPN [108] and 2,048 × 512 for UperNet [109].

Method	Backbone	#P(M)	GFLOPs	mIoU (%)
Semantic FPN [108]	PVTv2-B0 [80]	8	25	37.2
	VAN-B0	8	26	**38.5**
	ResNet18 [5]	16	32	32.9
	PVT-Tiny [20]	17	33	35.7
	PoolFormer-S12 [98]	15	31	37.2
	PVTv2-B1 [80]	18	34	42.5
	VAN-B1	18	35	**42.9**
	ResNet50 [5]	29	46	36.7
	PVT-Small [20]	28	45	39.8
	PoolFormer-S24 [98]	23	39	40.3
	PVTv2-B2 [80]	29	46	45.2
	VAN-B2	30	48	**46.7**
	ResNet101 [5]	48	65	38.8
	ResNeXt101-32x4d [7]	47	65	39.7
	PVT-Medium [20]	48	61	43.5
	PoolFormer-S36 [98]	35	48	42.0
	PVTv2-B3 [80]	49	62	47.3
	VAN-B3	49	68	**48.1**
UperNet [109]	ResNet-101 [5]	86	1029	44.9
OCRNet [41]		56	923	45.3
HamNet [42]		69	1111	46.8
UperNet [109]	Swin-T [15]	60	945	46.1
	ConvNeXt-T [21]	60	939	46.7
	VAN-B2	57	948	**50.1**
	Swin-S [15]	81	1038	49.3
	ConvNeXt-S [21]	82	1027	49.5
	VAN-B3	75	1030	**50.6**
	Swin-B [15]	121	1188	49.7
	ConvNeXt-B [21]	122	1170	49.9
	VAN-B4	90	1098	**52.2**

TABLE 14
Experimental results on COCO panoptic segmentation. * means model is pretrained on ImageNet-22K dataset. All methods are based on Mask2Former [113]. PQ means panoptic quality.

Backbone	Query type	Epochs	PQ	PQ^{Th}	PQ^{St}
Swin-T	100 queries	50	53.2	59.3	44.0
VAN-B2	100 queries	50	54.9	61.2	45.3
Swin-L*	200 queries	50	57.8	64.2	48.1
VAN-B6*	200 queries	50	**58.2**	**64.8**	**48.2**

Pose Estimation

TABLE 15
Comparison with the state-of-the-art vision backbones on COCO benchmark for pose estimation. Models are based SimpleBaseline [114].

Backbone	Input size	AP	AP^{50}	AP^{75}	AR	#P (M)	GFLOPs
HRNet-W32 [18]	256 × 192	74.4	90.5	81.9	78.9	28.5	7.1
PVT-S [20]	256 × 192	71.4	89.6	79.4	77.3	28.2	4.1
Swin-T [15]	256 × 192	72.4	90.1	80.6	78.2	32.8	6.1
Swin-B [15]	256 × 192	72.9	89.9	80.8	78.6	93.2	18.6
VAN-B2	256 × 192	**74.9**	**90.8**	**82.5**	**80.3**	30.3	6.1
HRNet-W32 [18]	384 × 288	75.8	90.6	82.7	81.0	28.5	16.0
Swin-B [15]	384 × 288	74.9	90.5	81.8	80.3	93.2	39.2
VAN-B2 [15]	384 × 288	**76.7**	**91.0**	**83.1**	**81.7**	30.3	13.6

Fine-grain Classification

TABLE 16
Experimental results on CUB-200 fine-grain classification dataset. * means model is pretrained on ImageNet-22K dataset.

Method	Backbone	Top-1 Acc (%)
ResNet-50 [5]	ResNet-101	84.5
ViT [13]	ViT-B_16*	90.3
DeiT [19]	DeiT-B*	90.0
VAN	VAN-B4*	91.3

TABLE 17
Comparing with different backbones on saliency detection task.

Backbone	DUTS-TE		DUT-O		PASCAL-S	
	F_{max}	MAE	F_{max}	MAE	F_{max}	MAE
ResNet18 [5]	0.853	0.044	0.769	0.056	0.854	0.071
PVT-T [20]	0.876	0.039	0.813	0.052	0.868	0.067
VAN-B1	**0.912**	**0.030**	**0.835**	**0.046**	**0.893**	**0.055**
ResNet50 [5]	0.873	0.038	0.786	0.051	0.864	0.065
PVT-S [20]	0.900	0.032	0.832	0.050	0.883	0.060
VAN-B2	**0.919**	**0.028**	**0.844**	**0.045**	**0.897**	**0.053**

THANKS