西湖大學 WESTLAKE UNIVERSITY

浙江大學 ZHEJIANG UNIVERSITY

ICLR
The 12th Internaltional Conference on Learning Representations

MogaNet: Efficient Multi-order Gated Aggregation Network

Siyuan Li*, Zedong Wang*, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, Stan Z. Li#

Westlake University, Zhejiang University

Timeline of Modern CNNs

InceptionNeXt CVPR'2024 (2023)

DCN.V4 CVPR'2024

ConvNeXt CVPR'2022

RepLKNet CVPR'2022

SLaK ICLR'2023

ConvNeXt.V2 CVPR'2023

DCN.V3 CVPR'2023

UniRepLKNet CVPR'2024

Convolution Kernel Designs

Large-Kernel Convolutions + Gated Attentions

VAN (2022) CVMJ'2023

HorNet NeurIPS'2022

FocalNet NeurIPS'2022

MogaNet (2022) ICLR'2024

Mamba arXiv'2023

VMamba arXiv'2024

Content

The 12th Internaltional Conference on Learning Representations

1. **Background: Macro Design and Pre-training**

   MetaFormer, ConvNeXt, ConvNeXt.V2 (SparK, A2MIM)

2. **Background: Design of Convolution Kernels**

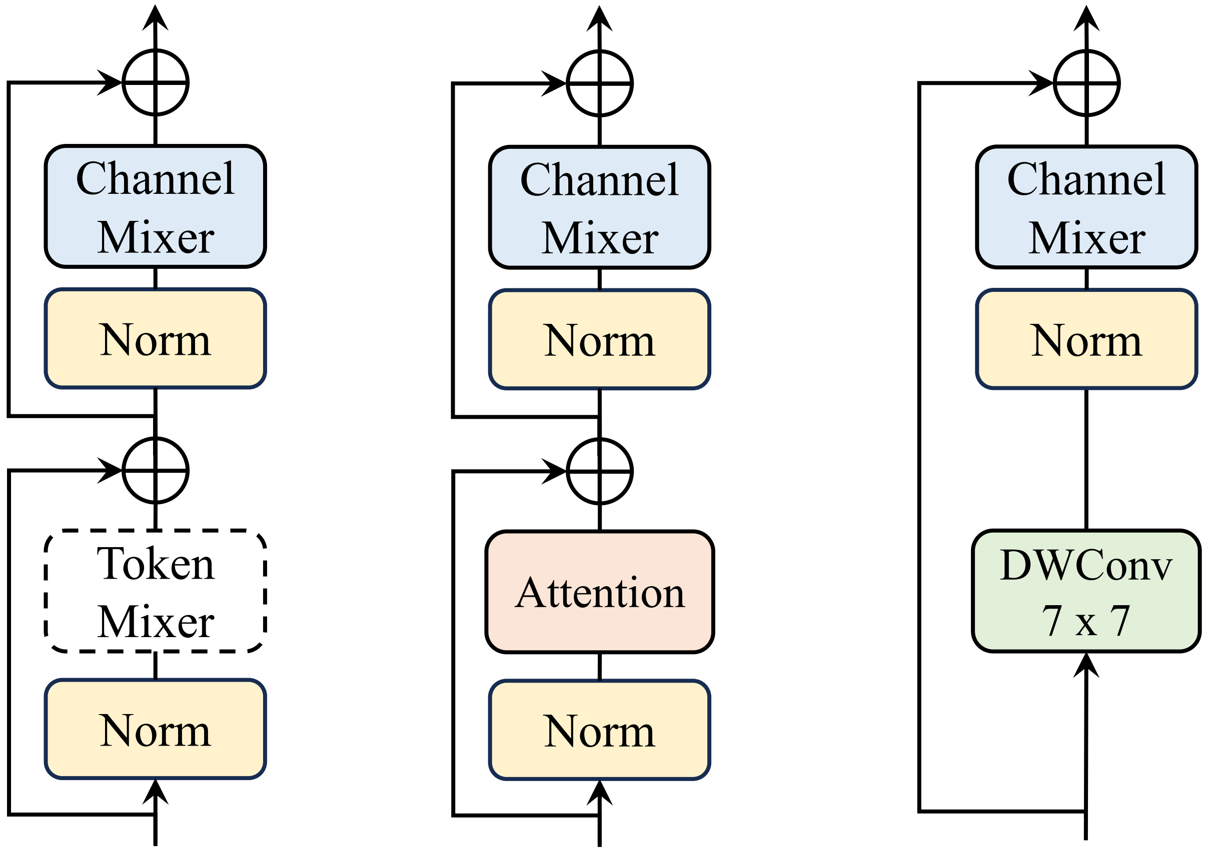
   RepLKNet, SLaK, InceptionNext, DCN.V3/V4, UniRepLKNet

3. **Combining Large Kernel with Gated Attention**

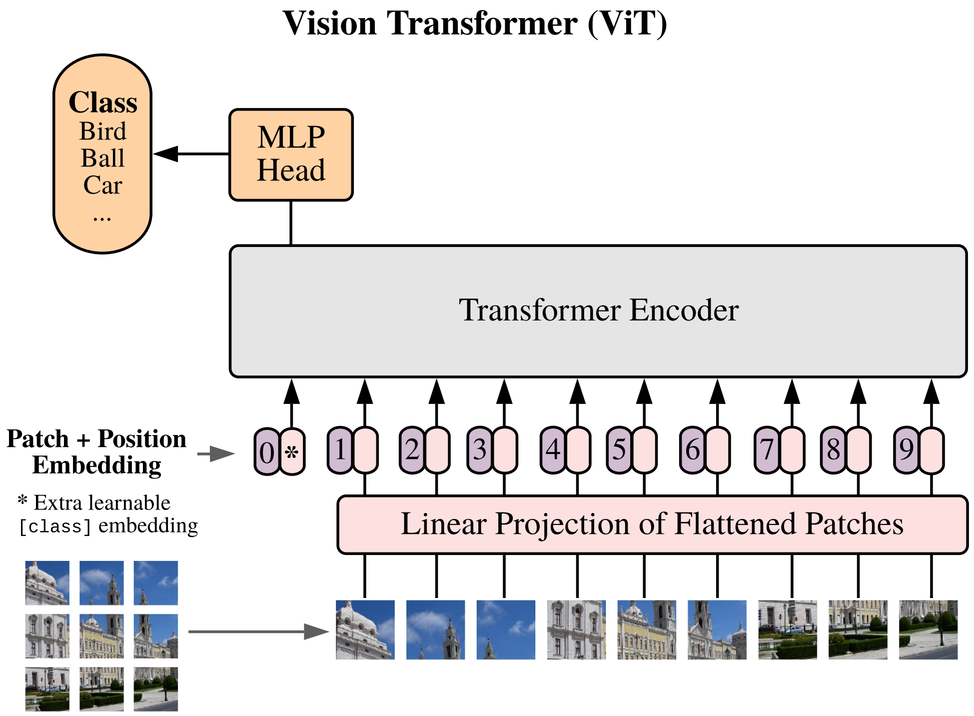
   VAN, HorNet, FocalNet, MogaNet, Mamba, VMamba, etc.

Modern CNNs: Macro Design

- Macro Design: Patch Embedding + Token Mixing + Channel Mixing + Pre-Norm & Short-cut.

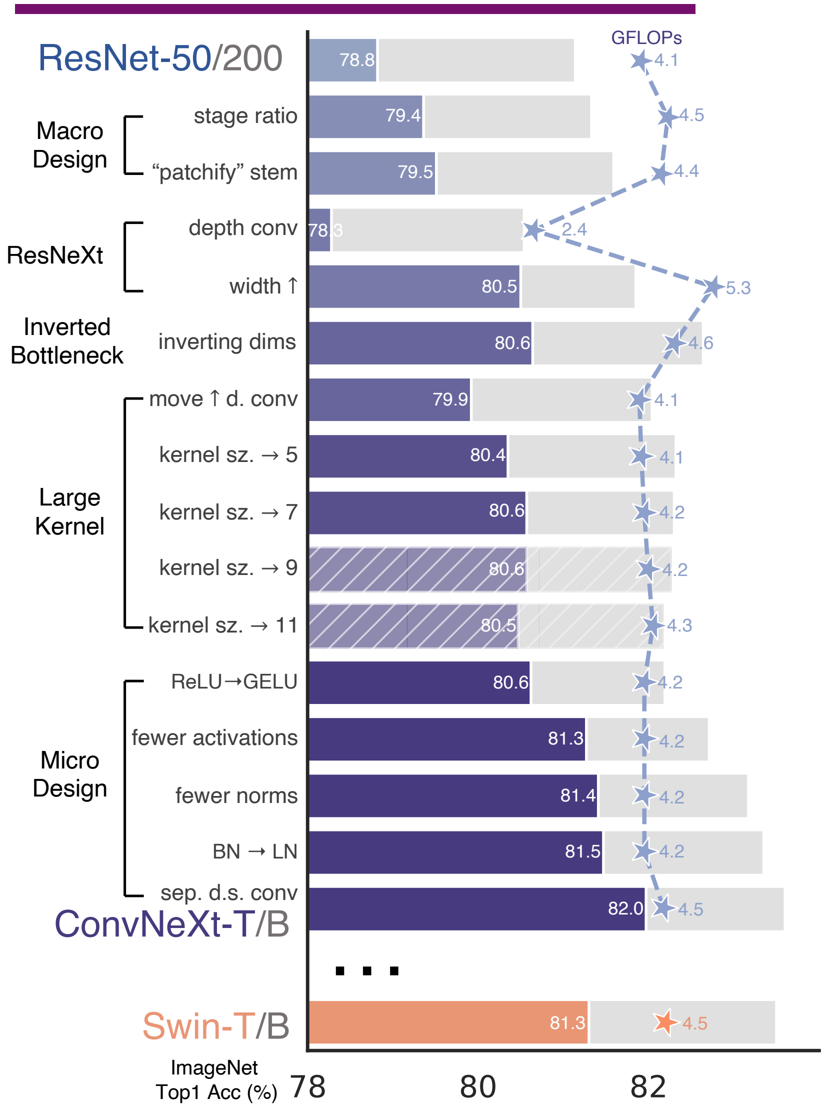


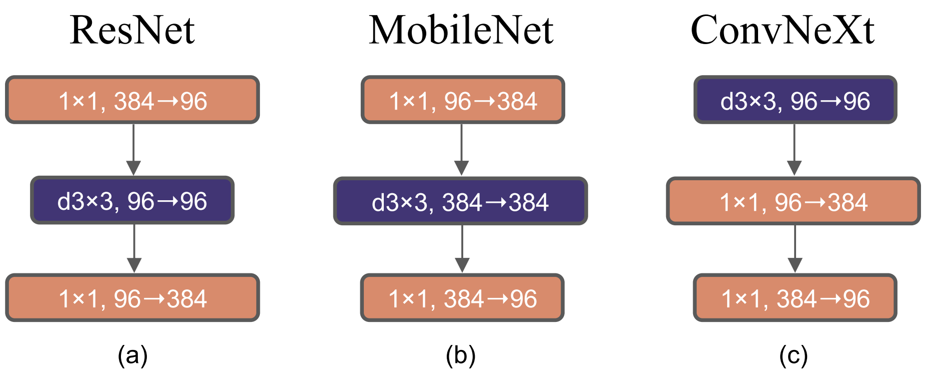
MetaFormer TransFormer ConvNeXt

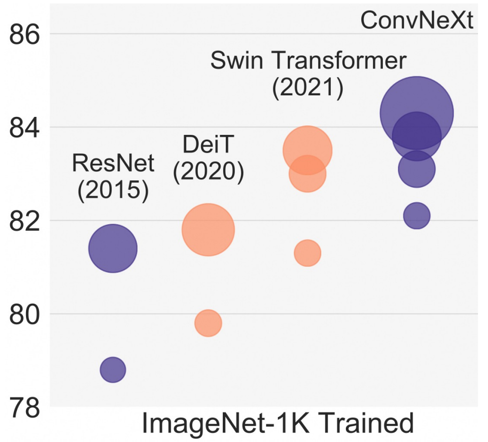


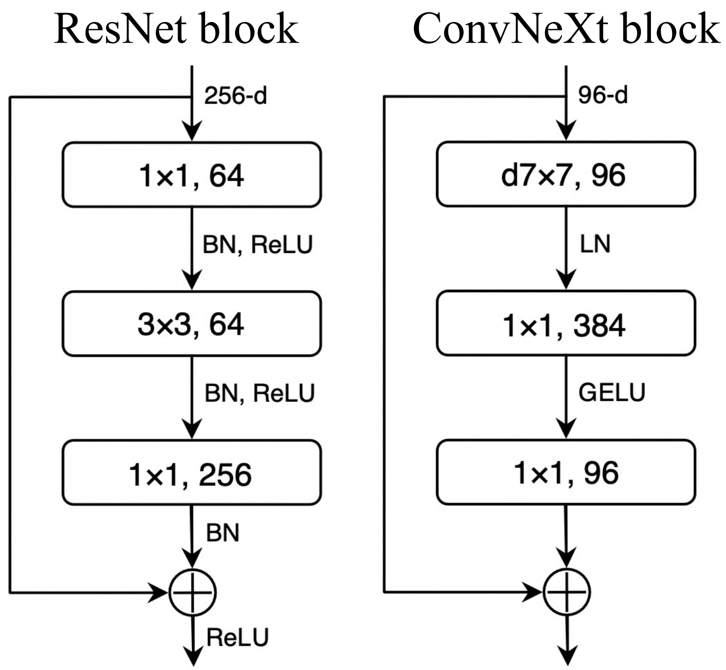
[1] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021
[2] PoolFormer: MetaFormer Is Actually What You Need for Vision. CVPR, 2022.
[3] A ConvNet for the 2020s. CVPR, 2022.

Modern CNNs: ConvNeXt







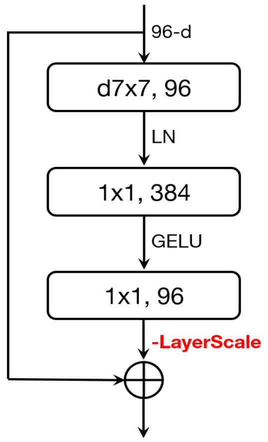


model	image size	#param.	FLOPs	throughput (image / s)	IN-1K top-1 acc.
ImageNet-1K trained models					
• RegNetY-16G [54]	224^2	84M	16.0G	334.7	82.9
• EffNet-B7 [71]	600^2	66M	37.0G	55.1	84.3
• EffNetV2-L [72]	480^2	120M	53.0G	83.7	85.7
○ DeiT-S [73]	224^2	22M	4.6G	978.5	79.8
○ DeiT-B [73]	224^2	87M	17.6G	302.1	81.8
○ Swin-T	224^2	28M	4.5G	757.9	81.3
• ConvNeXt-T	224^2	29M	4.5G	774.7	**82.1**
○ Swin-S	224^2	50M	8.7G	436.7	83.0
• ConvNeXt-S	224^2	50M	8.7G	447.1	**83.1**
○ Swin-B	224^2	88M	15.4G	286.6	83.5
• ConvNeXt-B	224^2	89M	15.4G	292.1	**83.8**
○ Swin-B	384^2	88M	47.1G	85.1	84.5
• ConvNeXt-B	384^2	89M	45.0G	95.7	**85.1**
• ConvNeXt-L	224^2	198M	34.4G	146.8	**84.3**
• ConvNeXt-L	384^2	198M	101.0G	50.4	**85.5**

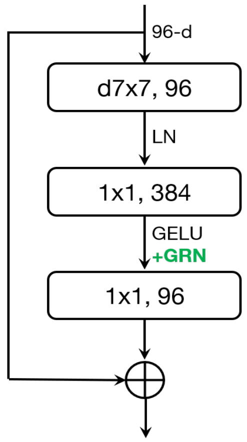
[1] A ConvNet for the 2020s. CVPR, 2022.

Modern CNNs: ConvNeXt.V2

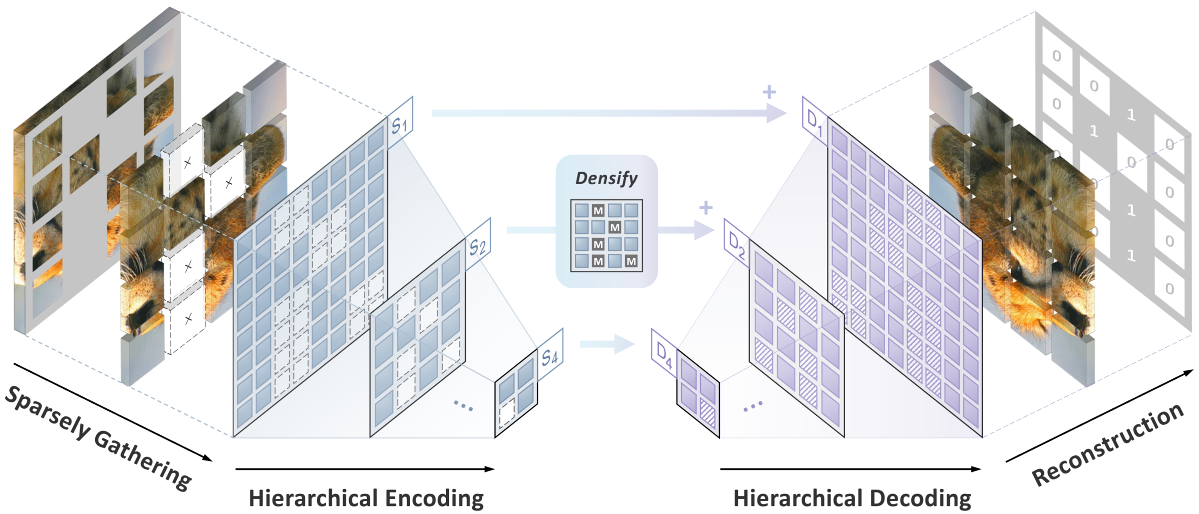
- **CNNs benefit from Masked Image Modeling (MIM) Pre-training.**



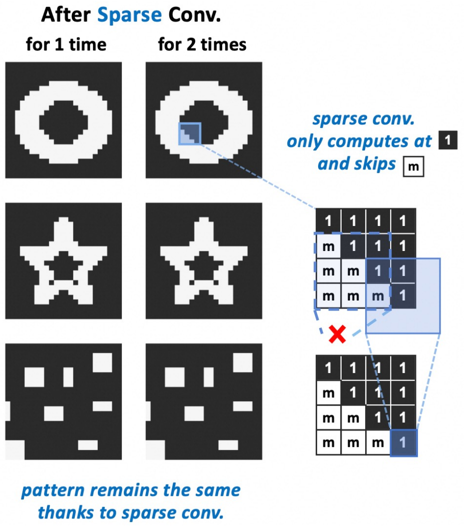
ConvNeXt.V1



ConvNeXt.V2



MIM pre-training with SparK (or FCMAE in ConvNeXt.V2)



Sparse Conv for Masking

Global Response Normalization (GRN)

```
# gamma, beta: learnable affine transform parameters
# X: input of shape (N,H,W,C)

gx = torch.norm(X, p=2, dim=(1,2), keepdim=True)
nx = gx / (gx.mean(dim=-1, keepdim=True)+1e-6)
return gamma * (X * nx) + beta + X
```

$$\mathcal{G}(X) := X \in \mathcal{R}^{H\times W\times C} \rightarrow gx \in \mathcal{R}^{C}$$

$$\mathcal{N}(||X_i||) := ||X_i|| \in \mathcal{R} \rightarrow \frac{||X_i||}{\sum_{j=1,\dots,C}||X_j||} \in \mathcal{R}$$

Backbone	Method	#param	FLOPs	Val acc.
ConvNeXt V1-B	Supervised	89M	15.4G	83.8
ConvNeXt V1-B	FCMAE	89M	15.4G	83.7
ConvNeXt V2-B	Supervised	89M	15.4G	84.3 (+0.5)
ConvNeXt V2-B	FCMAE	89M	15.4G	**84.6** (+0.8)
ConvNeXt V1-L	Supervised	198M	34.4G	84.3
ConvNeXt V1-L	FCMAE	198M	34.4G	84.4
ConvNeXt V2-L	Supervised	198M	34.4G	84.5 (+0.2)
ConvNeXt V2-L	FCMAE	198M	34.4G	**85.6** (+1.3)

Methods Target	#Para. (M)	Sup. Label	MoCoV3‡ CL	SimMIM‡ RGB	SparK RGB	A^{2}MIM RGB
ResNet-50	25.6	79.8	80.1	79.9	80.6	**80.4**
ResNet-101	44.5	81.3	81.6	81.3	82.2	**81.9**
ResNet-152	60.2	81.8	82.0	81.9	82.7	**82.5**
ResNet-200	64.7	82.1	82.5	82.2	83.1	**83.0**
ConvNeXt-T	28.6	82.1	82.3	82.1	82.7	**82.5**
ConvNeXt-S	50.2	83.1	83.3	83.2	84.1	**83.7**
ConvNeXt-B	88.6	83.5	83.7	83.6	84.8	**84.1**

[1] ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. CVPR, 2023. [2] Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling. ICLR, 2023. [3] Architecture-Agnostic Masked Image Modeling - From ViT back to CNN. ICML, 2023.

Content

1. Background: Macro Design and Pre-training

   MetaFormer, ConvNeXt, ConvNeXt.V2 (SparK, A2MIM)

2. Background: Design of Convolution Kernels

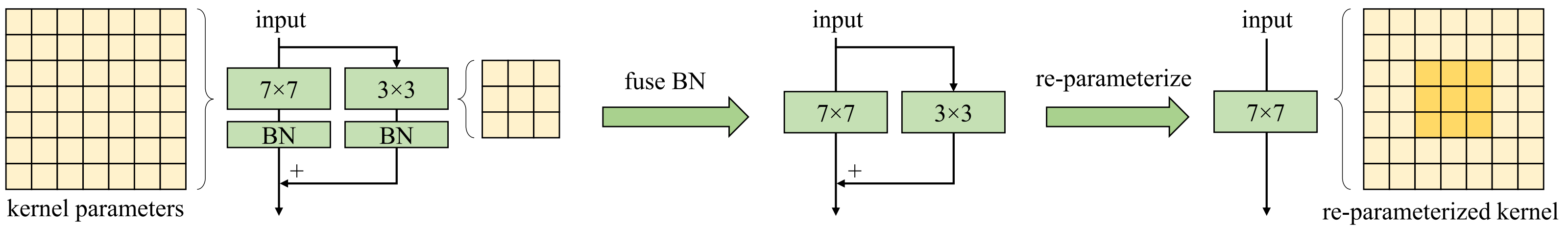
   RepLKNet, SLaK, InceptionNext, DCN.V3/V4, UniRepLKNet

3. Combining Large Kernel with Gated Attention

   VAN, HorNet, FocalNet, MogaNet, Mamba, VMamba, etc.

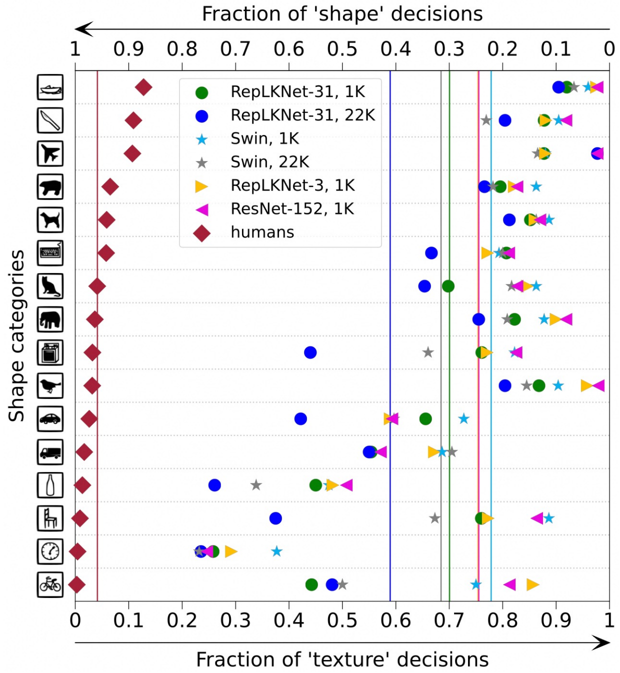
Large Kernels: RepLKNet

- Large-Kernel (LK) Convolutions are **efficient** and **competitive** as Self-attention.
- Training extremely large convolutions with **Structural Re-parameterization.**



DW7×7 = DW3×3 (BN) +DW7×7 (BN)+Short-cut.

Large kernels are **shape biased** as ViTs.

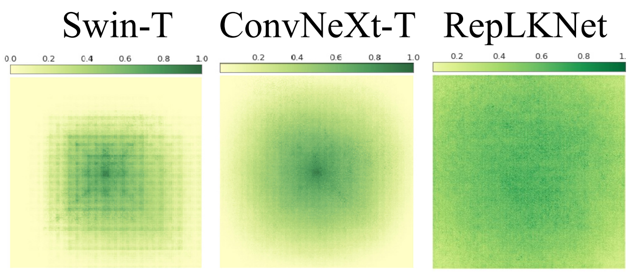


Resolution R	Impl	Latency (ms) @ Kernel size 3	5	7	9	13	17	21	27	29	31
16 × 16	Pytorch	5.6	11.0	14.4	17.6	36.0	57.2	83.4	133.5	150.7	171.4
	Ours	5.6	6.5	6.4	6.9	7.5	8.4	8.4	8.4	8.3	8.4
32 × 32	Pytorch	21.9	34.1	54.8	76.1	141.2	230.5	342.3	557.8	638.6	734.8
	Ours	21.9	28.7	34.6	40.6	52.5	64.5	73.9	87.9	92.7	96.7
64 × 64	Pytorch	69.6	141.2	228.6	319.8	600.0	977.7	1454.4	2371.1	2698.4	3090.4
	Ours	69.6	112.6	130.7	152.6	199.7	251.5	301.0	378.2	406.0	431.7

Large kernels are **memory bound** instead of compute bound.

Kernel size	Architecture	ImageNet Top-1	Params	FLOPs	ADE20K mIoU	Params	FLOPs
7-7-7-7	ConvNeXt-Tiny	81.0	29M	4.5G	44.6	60M	939G
7-7-7-7	ConvNeXt-Small	82.1	50M	8.7G	45.9	82M	1027G
7-7-7-7	ConvNeXt-Base	82.8	89M	15.4G	47.2	122M	1170G
31-29-27-13	ConvNeXt-Tiny	81.6	32M	6.1G	**46.2**	64M	973G
31-29-27-13	ConvNeXt-Small	82.5	58M	11.3G	**48.2**	90M	1081G

Extremely large kernels benefit both classification and downstream tasks and outperforms ViTs.

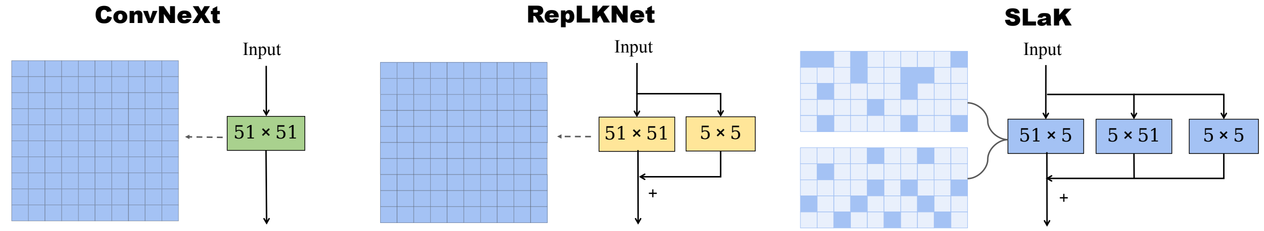


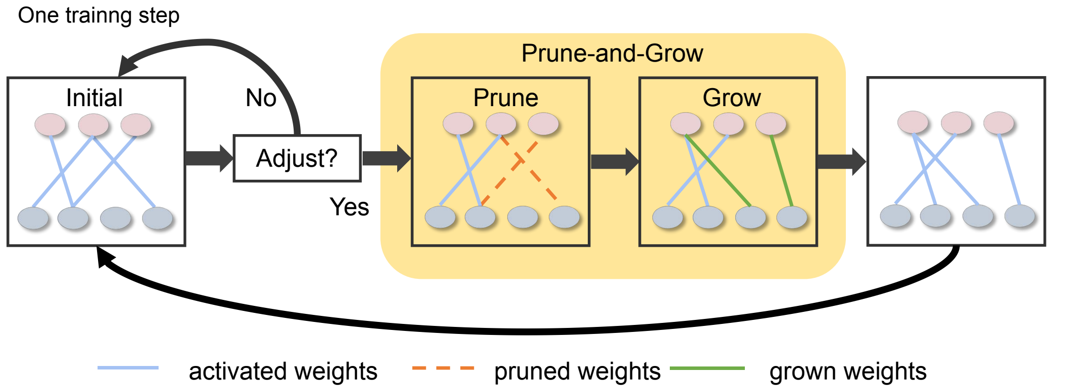
Effective receptive field

[1] Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs. CVPR, 2022.

Large Kernels: SLaK

- Step 1: Decomposing a large kernel (61x61) into two rectangular, parallel kernels.
- Step 2: Using sparse groups training (speedup), expanding more width.

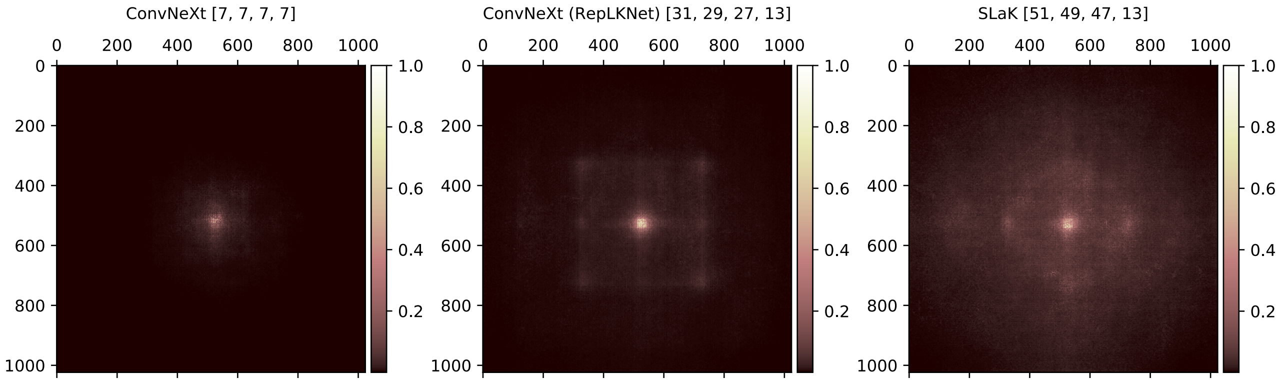




(1) Initialization: Constructing Sparce Convolution based on SNIP[2]
(2) Dynamic sparsity: Pruning (the lowest magnitude) and growing

Kernel Size	Top-1 Acc	#Params	FLOPs	Top-1 Acc	#Params	FLOPs	Top-1 Acc	#Params	FLOPs
	Decomposed			Sparse groups			Sparse groups, expand more width		
7-7-7-7	81.0	29M	4.5G	80.0	17M	2.6G	81.1	29M	4.5G
31-29-37-13	81.3	30M	5.0G	80.4	18M	2.9G	81.5	30M	4.8G
51-49-47-13	81.5	31M	5.4G	80.5	18M	3.1G	81.6	30M	5.0G
61-59-57-13	81.4	31M	5.6G	80.4	19M	3.2G	81.5	31M	5.2G

Model	Kernel Size	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
pre-trained for 120 epochs, finetuned for 1× (12 epochs)							
ConvNeXt-T (Liu et al., 2022b)	7-7-7-7	47.3	65.9	51.5	41.1	63.2	44.4
ConvNeXt-T (RepLKNET)* (Ding et al., 2022)	31-29-27-13	47.8	66.7	52.0	41.4	63.9	44.7
SLaK-T	51-49-47-13	**48.4**	**67.2**	**52.5**	**41.8**	**64.4**	**45.2**
pre-trained for 300 epochs, finetuned for 3× (36 epochs)							
ConvNeXt-T (Liu et al., 2022b)	7-7-7-7	50.4	69.1	54.8	43.7	66.5	47.3
SLaK-T	51-49-47-13	**51.3**	**70.0**	**55.7**	**44.3**	**67.2**	**48.1**



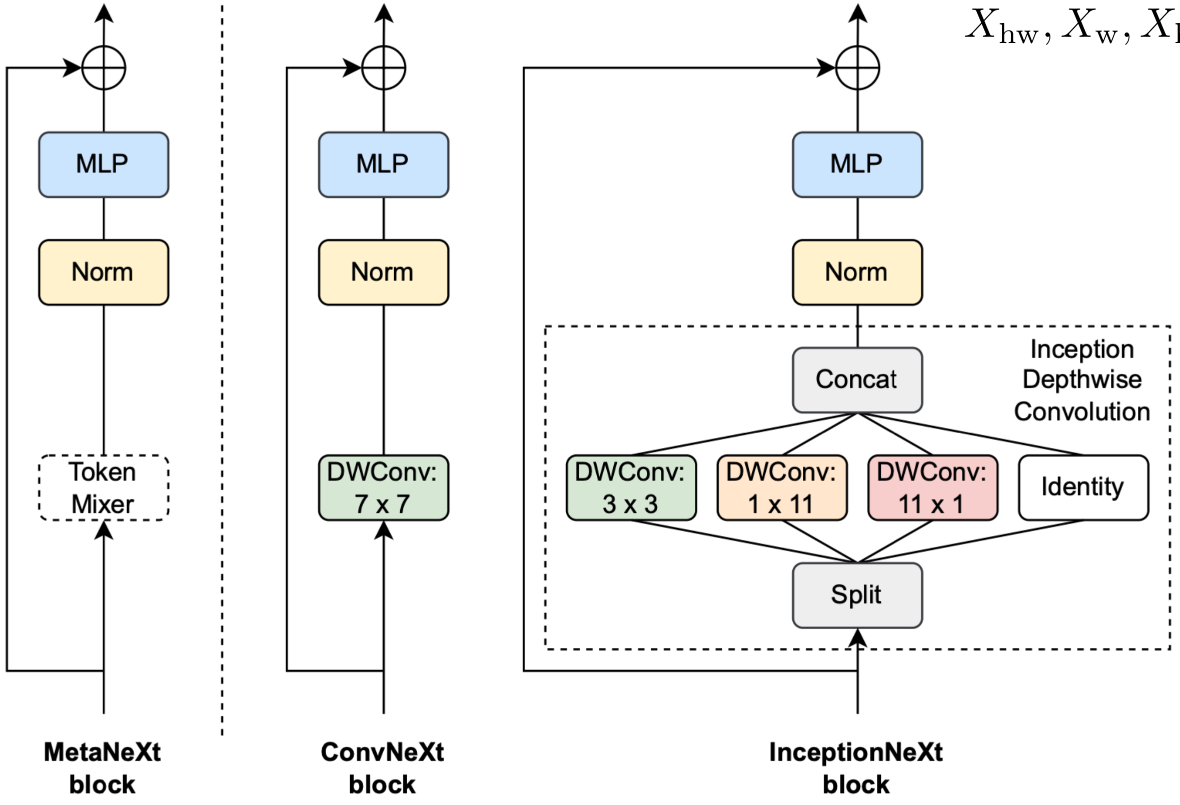
Effective receptive field (ERF)

[1] More ConvNets in the 2020s: Scaling up Kernels Beyond 51x51 using Sparsity. ICLR, 2023.
[2] SNIP: Single-shot Network Pruning based on Connection Sensitivity. ICLR, 2019.

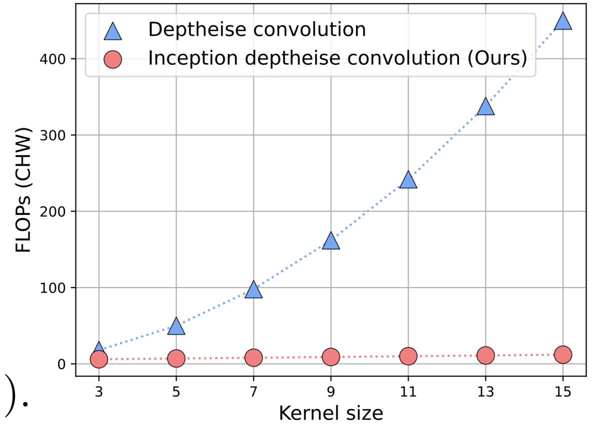
Large Kernels: InceptionNeXt

The 12th Internaltional Conference on Learning Representations

- MetaNeXt: Fusing Token Mixer with Channel Mixer + PreNorm + ShortCut.
- Inception Kernels: Better performance and throughputs than Depth-wise Conv 7x7.



$$
\begin{aligned}
X_{\mathrm{hw}}, X_{\mathrm{w}}, X_{\mathrm{h}}, X_{\mathrm{id}} &= \mathrm{Split}(X) \\
&= X_{:,:g}, X_{:g:2g}, X_{:2g:3g}, X_{:3g:} \\
X'_{\mathrm{hw}} &= \mathrm{DWConv}^{g \to g}_{k_s \times k_s} g(X_{\mathrm{hw}}), \\
X'_{\mathrm{w}} &= \mathrm{DWConv}^{g \to g}_{1 \times k_b} g(X_{\mathrm{w}}), \\
X'_{\mathrm{h}} &= \mathrm{DWConv}^{g \to g}_{k_b \times 1} g(X_{\mathrm{h}}), \\
X'_{\mathrm{id}} &= X_{\mathrm{id}}. \\
X' &= \mathrm{Concat}(X'_{\mathrm{hw}}, X'_{\mathrm{w}}, X'_{\mathrm{h}}, X'_{\mathrm{id}}).
\end{aligned}
$$

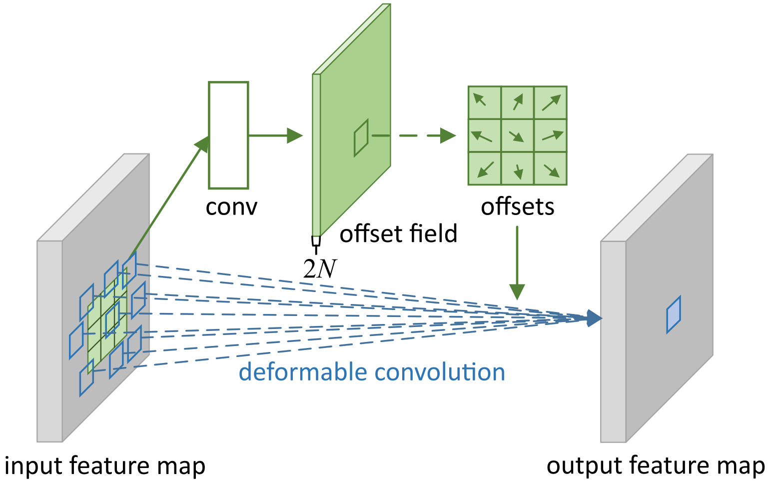


Model	Mixing Type	Image (size)	Params (M)	MACs (G)	Throughput (img/second) Train	Inference	Top-1 (%)
DeiT-S [61]	Attn	224^2	22	4.6	1227	3781	79.8
T2T-ViT-14 [76]	Attn	224^2	22	4.8	–	–	81.5
TNT-S [18]	Attn	224^2	24	5.2	–	–	81.5
Swin-T [37]	Attn	224^2	29	4.5	564	1768	81.3
Focal-T [73]	Attn	224^2	29	4.9	–	–	82.2
ResNet-50 [20, 69]	Conv	224^2	26	4.1	969	3149	78.4
RSB-ResNet-50 [20, 69]	Conv	224^2	26	4.1	969	3149	79.8
RegNetY-4G [46, 69]	Conv	224^2	21	4.0	670	2694	81.3
FocalNet-T [72]	Conv	224^2	29	4.5	–	–	82.3
ConvNeXt-T [38]	Conv	224^2	29	4.5	575	2413 (1943)	82.1
InceptionNeXt-T (Ours)	Conv	224^2	28	4.2	901 (+57%)	2900 (+20%)	82.3 (+0.2)

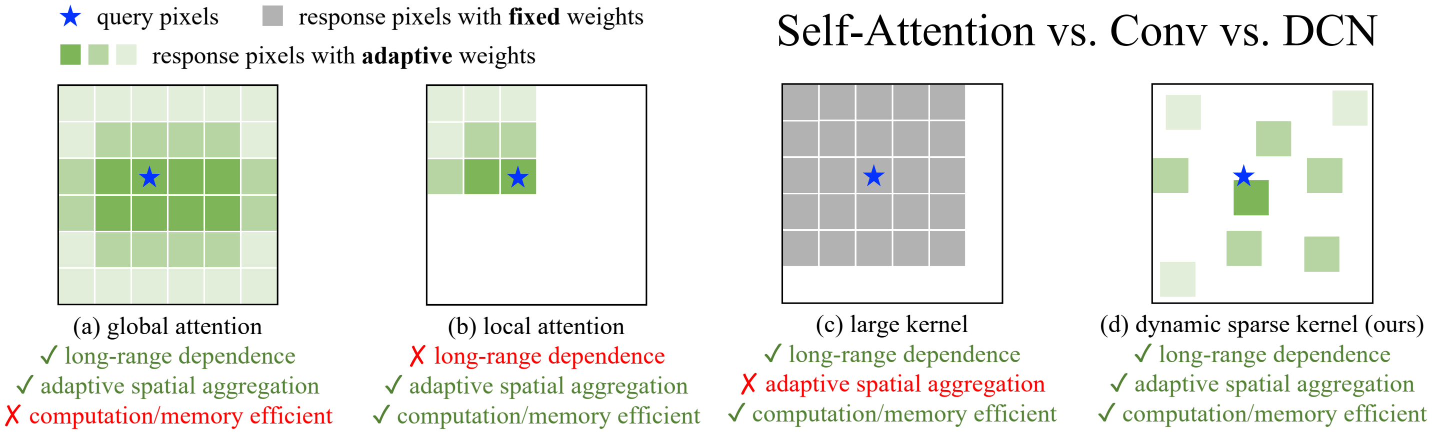
[1] InceptionNeXt: When Inception Meets ConvNeXt. CVPR, 2024.

Kernel Designs: DCN.V3 (InternImage)

- DCN.V3: Learnable offsets (V1) + Softmax-normalized modulation (V2) + Grouping.



Self-Attention vs. Conv vs. DCN



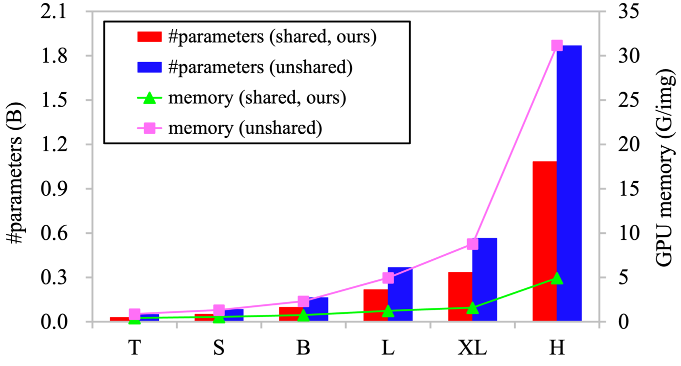
DCN.V1: $\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta \mathbf{p}_n)$

DCN.V2: $\mathbf{y}(p_0) = \sum_{k=1}^{K} \mathbf{w}_k \mathbf{m}_k \mathbf{x}(p_0 + p_k + \Delta p_k)$

DCN.V3: $\mathbf{y}(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} \mathbf{w}_g \mathbf{m}_{gk} \mathbf{x}_g(p_0 + p_k + \Delta p_{gk})$

Offsets Δp_n, Regular grids p_n, Modulation m_k, weights w

Scaling-up with efficient impl.

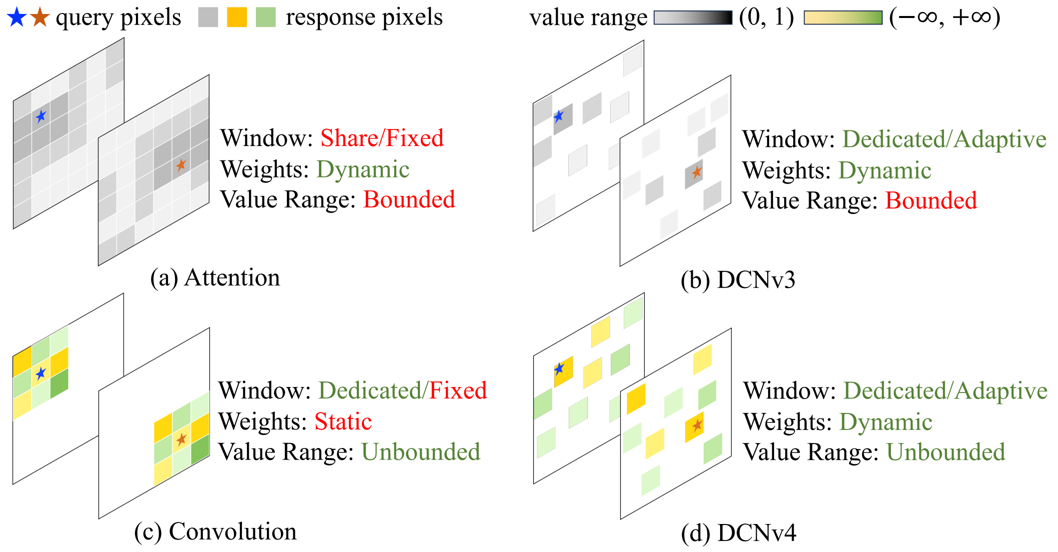


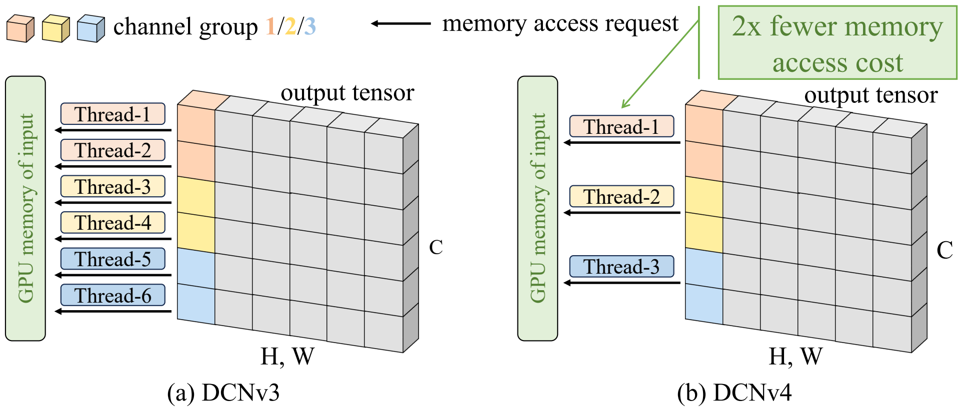
method	type	scale	#params	#FLOPs	acc (%)
SwinV2-L/24‡ [16]	T	384^2	197M	115G	87.6
RepLKNet-31L‡ [22]	C	384^2	172M	96G	86.6
HorNet-L‡ [43]	C	384^2	202M	102G	87.7
ConvNeXt-L‡ [21]	C	384^2	198M	101G	87.5
ConvNeXt-XL‡ [21]	C	384^2	350M	179G	87.8
InternImage-L‡ (ours)	C	384^2	223M	108G	87.7
InternImage-XL‡ (ours)	C	384^2	335M	163G	88.0
ViT-G/14# [30]	T	518^2	1.84B	5160G	90.5
CoAtNet-6# [20]	T	512^2	1.47B	1521G	90.5
CoAtNet-7# [20]	T	512^2	2.44B	2586G	90.9
Florence-CoSwin-H# [59]	T	–	893M	–	90.0
SwinV2-G# [16]	T	640^2	3.00B	–	90.2
RepLKNet-XL# [22]	C	384^2	335M	129G	87.8
BiT-L-ResNet152x4# [67]	C	480^2	928M	–	87.5
InternImage-H# (ours)	C	224^2	1.08B	188G	88.9
InternImage-H# (ours)	C	640^2	1.08B	1478G	89.6

[1] Deformable Convolutional Networks. ICCV, 2017.
[2] Deformable ConvNets v2: More Deformable, Better Results. CVPR, 2018.
[3] InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. CVPR, 2023.

Kernel Designs: DCN.V4

- DCN.V4: No Softmax normalization + Speed-up (reducing HRM as Flash-Attention).





Model	5th EP	10th Ep	20th Ep	50th Ep	300th Ep
ConvNeXt	29.9	53.5	66.1	74.8	83.8
ConvNeXt + softmax	8.5 (-21.4)	25.3 (-28.2)	51.1 (-15.0)	69.1 (-5.7)	81.6 (-2.2)

Using Softmax in DWConv7×7 degenerating performance

Operator	Runtime (ms) 56 × 56 × 128	28 × 28 × 256	14 × 14 × 512	7 × 7 × 1024	14 × 14 × 768
Attention (torch)	30.8 / 19.3	3.35 / 2.12	0.539 / 0.448	0.446 / 0.121	0.779 / 0.654
FlashAttention-2	N/A / 2.46	N/A / 0.451	N/A / **0.123**	N/A / 0.0901	N/A / 0.163
Window Attn (7 × 7)	4.05 / 1.46	2.07 / 0.770	1.08 / 0.422	0.577 / 0.239	1.58 / 0.604
DWConv (7 × 7, torch)	2.02 / 1.98	1.03 / 1.00	0.515 / 0.523	0.269 / 0.261	0.779 / 0.773
DWConv (7 × 7, cuDNN)	0.981 / 0.438	0.522 / 0.267	0.287 / 0.153	0.199 / 0.102	0.413 / 0.210
DCNv3	1.45 / 1.52	0.688 / 0.711	0.294 / 0.298	0.125 / 0.126	0.528 / 0.548
DCNv4	**0.606 / 0.404**	**0.303 / 0.230**	**0.145 / 0.123**	**0.0730 / 0.0680**	**0.224 / 0.147**

ImageNet-1K Classification

Model	Size	Scale	Acc	Throughput
Swin-T	29M	224^2	81.3	1989 / 3619
ConvNeXt-T	29M	224^2	82.1	2485 / 4305
InternImage-T	30M	224^2	83.5	1409 / 1746
FlashInternImage-T	30M	224^2	**83.6**	2316 / 3154 (+64%/ + 80%)
Swin-S	50M	224^2	83.0	1167/2000
ConvNeXt-S	50M	224^2	83.1	1645/2538
InternImage-S	50M	224^2	84.2	1044/1321
FlashInternImage-S	50M	224^2	**84.4**	1625 / 2396

COCO2017 Det. and Seg.

Model	#param	FPS	Cascade Mask R-CNN 1× AP^b	1× AP^m	3×+MS AP^b	3×+MS AP^m
Swin-L	253M	20 / 26	51.8	44.9	53.9	46.7
ConvNeXt-L	255M	26 / 40	53.5	46.4	54.8	47.6
InternImage-L	277M	20 / 26	54.9	47.7	56.1	48.5
ConvNeXt-XL	407M	21 / 32	53.6	46.5	55.2	47.7
InternImage-XL	387M	16 / 23	55.3	48.1	56.2	48.8
FlashInternImage-L	277M	26 / 39	**55.6**	**48.2**	**56.7**	**48.9**

[1] DCNv4: Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications. CVPR, 2024.

Content

1. Background: Macro Design and Pre-training

   MetaFormer, ConvNeXt, ConvNeXt.V2 (SparK, A2MIM)

2. Background: Design of Convolution Kernels

   RepLKNet, SLaK, InceptionNext, DCN.V3/V4, UniRepLKNet

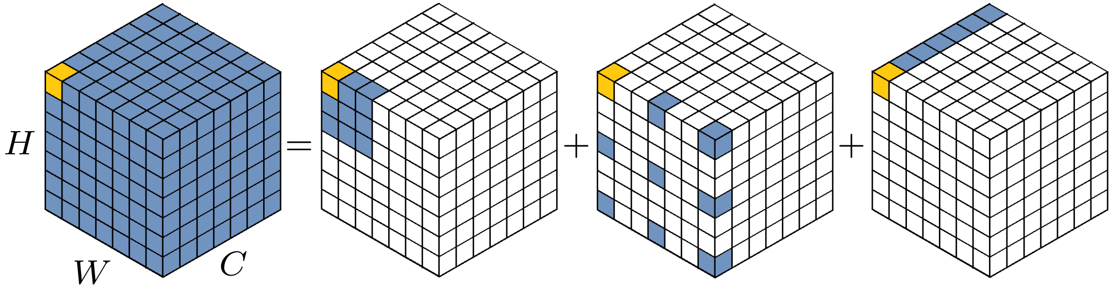
3. Combining Large Kernel with Gated Attention

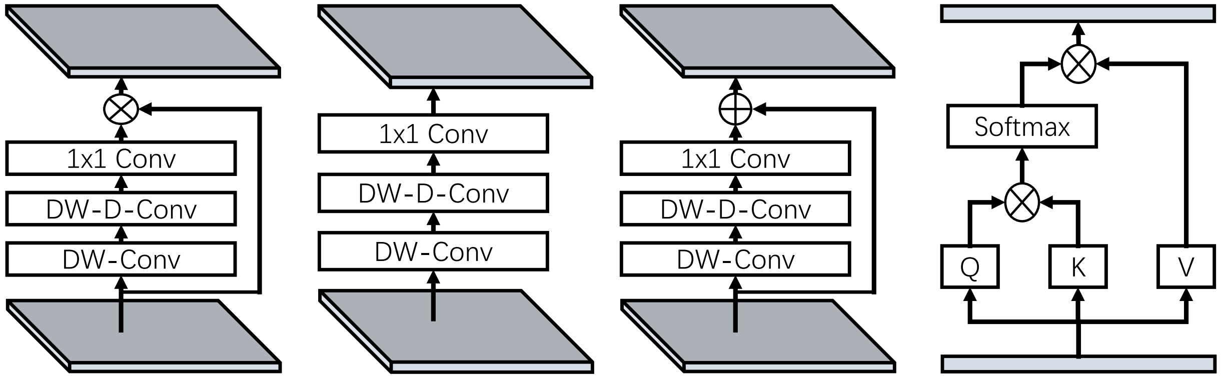
   VAN, HorNet, FocalNet, MogaNet, Mamba, VMamba, etc.

Gating & Large-kernel: VAN

- **Decomposed large kernel + Gating.**

Conv9×9 = DWConv3×3 + DWConv3×3 (Dilation=3) + PWConv1×1





VAN (LKA) Non-attention Non-attention (add) Self-attention

[1] Visual Attention Network. CVMJ, 2023.

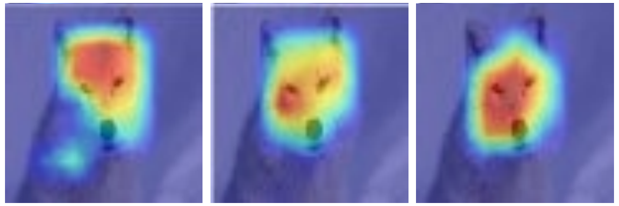
Properties	**Convolution**	**Self-Attention**	**LKA**
Local Receptive Field	✓	✗	✓
Long-range Dependence	✗	✓	✓
Spatial Adaptability	✗	✓	✓
Channel Adaptability	✗	✗	✓
Computational complexity	$\mathcal{O}(n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$

Properties of DWConv vs. MHSA vs. Large-kernel Attention

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

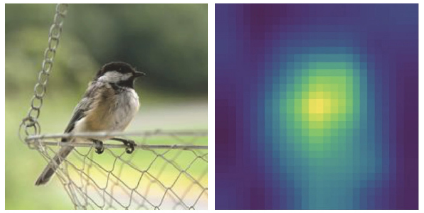
Kernel size vs. Dilation vs. ImageNet Acc (%)

Conv21×21 = DWConv5×5 +DWConv7×7 (Dilation=3) +PWConv1×1



Swin-T ConvNeXt-T VAN-B2

Grad-CAM visualization



Attention map visualization

MogaNet: Motivation

The 12th Internaltional Conference on Learning Representations

- **Representation Bottleneck[1]: Loss in the middle-order interactions.**

Multi-order Interactions $I^{(m)}(i,j) = \mathbb{E}_{S \subseteq N \setminus \{i,j\}, |S|=m}[\Delta f(i,j,S)]$

$N = \{1, \ldots, n\} \quad 0 \leq m \geq n-2$

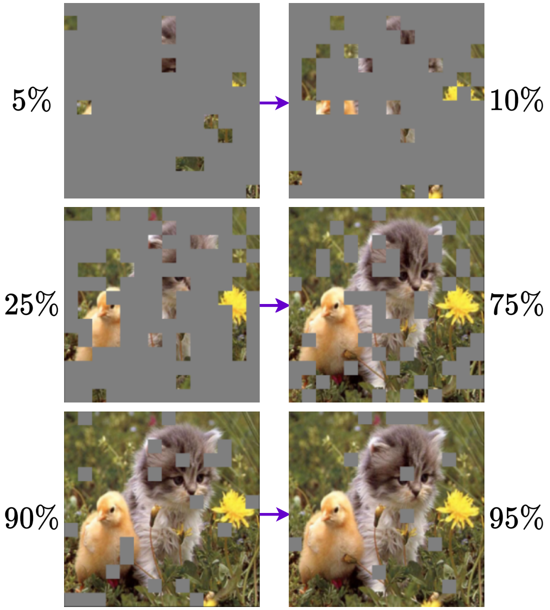
$\Delta f(i,j,S) = f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S)$

Interaction Strengths $J^{(m)} = \dfrac{\mathbb{E}_{x \in \Omega} \mathbb{E}_{i,j} |I^{(m)}(i,j|x)|}{\mathbb{E}_{m'} \mathbb{E}_{x \in \Omega} \mathbb{E}_{i,j} |I^{(m')}(i,j|x)|}$

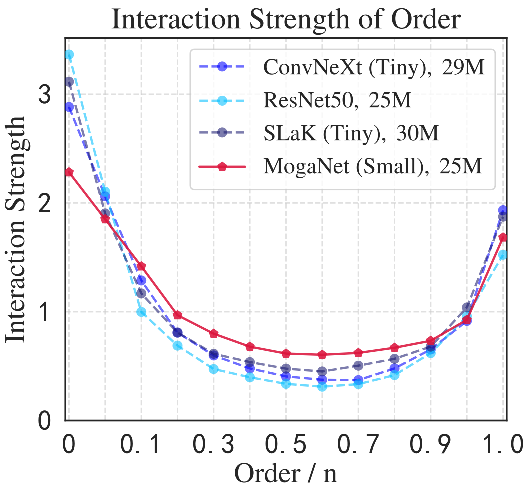
Much new information
Little new infomation

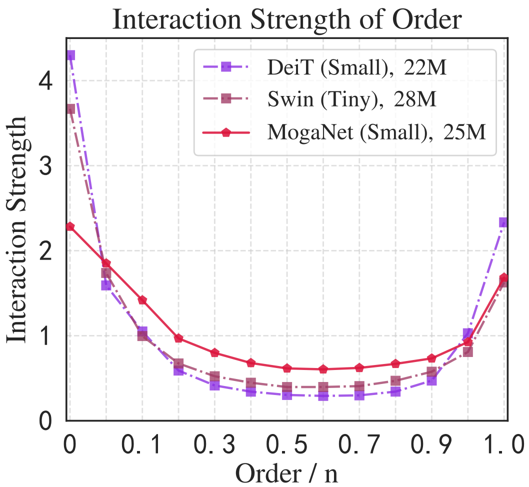
Little new information
Much new infomation

Much new information
Little new infomation



Both ViTs and modern CNN architectures fail to explore middle-order interactions, which are informative to humans.

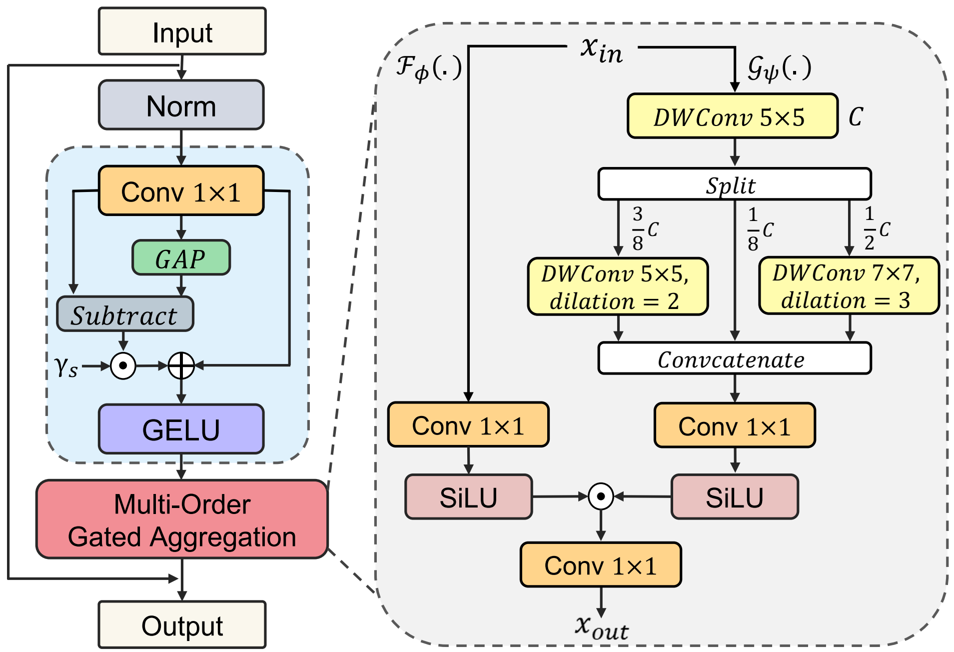




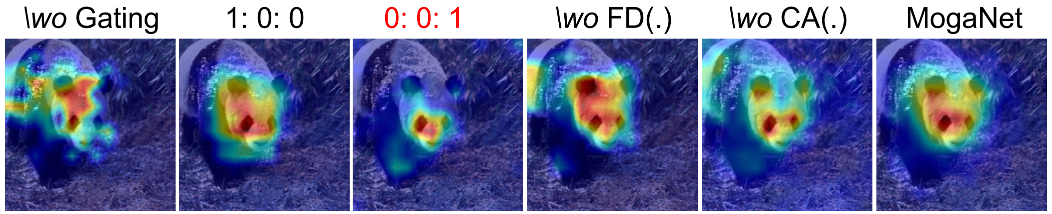
[1] Discovering and Explaining the Representation Bottleneck of DNNs. ICLR, 2022. [2] MogaNet: Multi-order Gated Aggregation Network. ICLR, 2024.

MogaNet: Spatial Aggregation

- Spatial Aggregation (SA): Multi-order context extraction + Gated aggregation.



$$Z = X + \mathrm{Moga}\Big(\mathrm{FD}\big(\mathrm{Norm}(X)\big)\Big)$$



Feature decomposition:

$$Y = \mathrm{Conv}_{1\times1}(X),$$
$$Z = \mathrm{GELU}\Big(Y + \gamma_s \odot \big(Y - \mathrm{GAP}(Y)\big)\Big)$$

Gated aggregation branch:

$$Z = \underbrace{\mathrm{SiLU}\big(\mathrm{Conv}_{1\times1}(X)\big)}_{\mathcal{F}_\phi} \odot \underbrace{\mathrm{SiLU}\big(\mathrm{Conv}_{1\times1}(Y_C)\big)}_{\mathcal{G}_\psi}$$

Multi-order DWConvs: DW5×5 (d=1), DW5×5 (d=2), DW7×7 (d=3)

$$C_l + C_m + C_h = C,\ Y_C = \mathrm{Concat}(Y_{l,1:C_l}, Y_m, Y_h)$$

Modules	Top-1 Acc (%)	Params. (M)	FLOPs (G)
Baseline (+Gating branch)	77.2	5.09	1.070
$\mathrm{DW}_{7\times7}$	77.4	5.14	1.094
$\mathrm{DW}_{5\times5,d=1} + \mathrm{DW}_{7\times7,d=3}$	77.5	5.15	1.112
$\mathrm{DW}_{5\times5,d=1} + \mathrm{DW}_{5\times5,d=2} + \mathrm{DW}_{7\times7,d=3}$	77.5	5.17	1.185
+Multi-order, $C_l : C_m : C_h = 1:0:3$	77.5	5.17	1.099
+Multi-order, $C_l : C_m : C_h = 0:1:1$	77.6	5.17	1.103
+Multi-order, $C_l : C_m : C_h = 1:6:9$	77.7	5.17	1.104
+Multi-order, $C_l : C_m : C_h = 1:3:4$	**77.8**	5.17	1.102

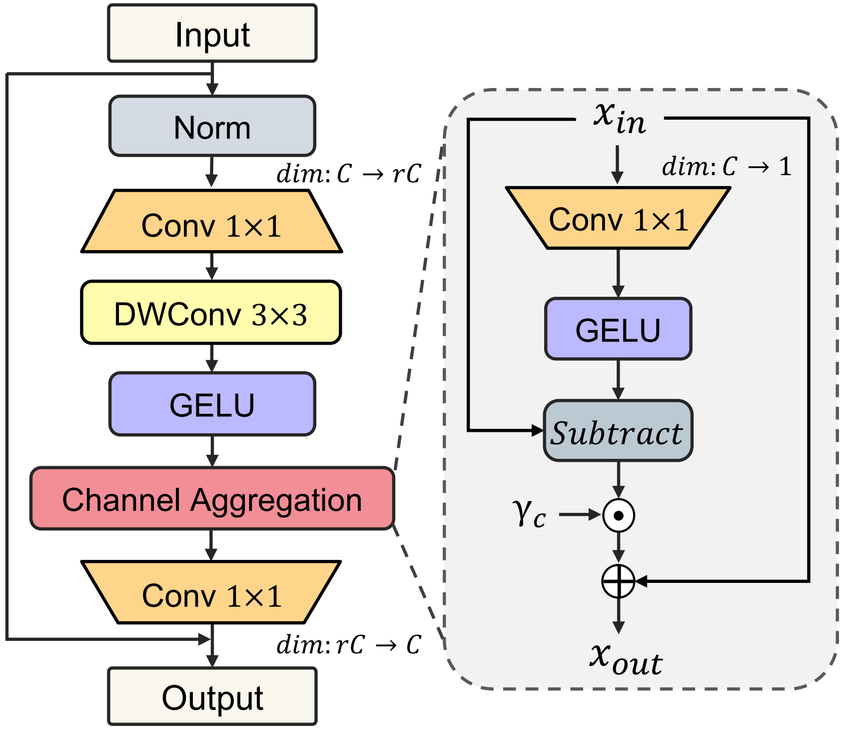
Top-1	Acc (%)	Context branch None	GELU	SiLU
Gating branch	None	76.3	76.7	76.7
	Sigmoid	76.8	77.0	76.9
	GELU	76.7	76.8	77.0
	SiLU	76.9	77.1	**77.2**

Ablation of SA module with MogaNet-T on ImageNet

[1] MogaNet: Multi-order Gated Aggregation Network. ICLR, 2024.

MogaNet: Channel Aggregation

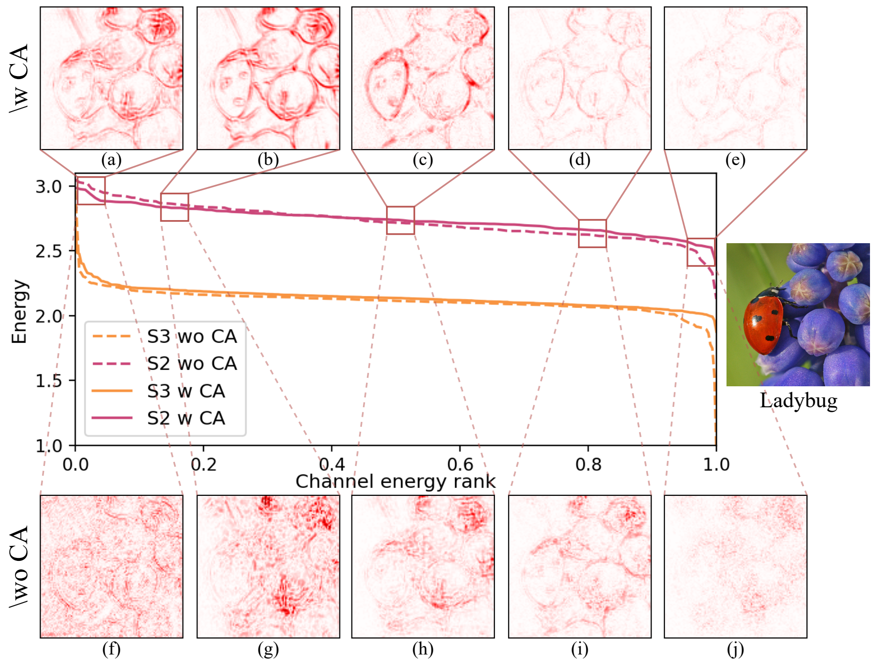
- **Channel Aggregation (CA): Multi-order Channel Reallocation.**



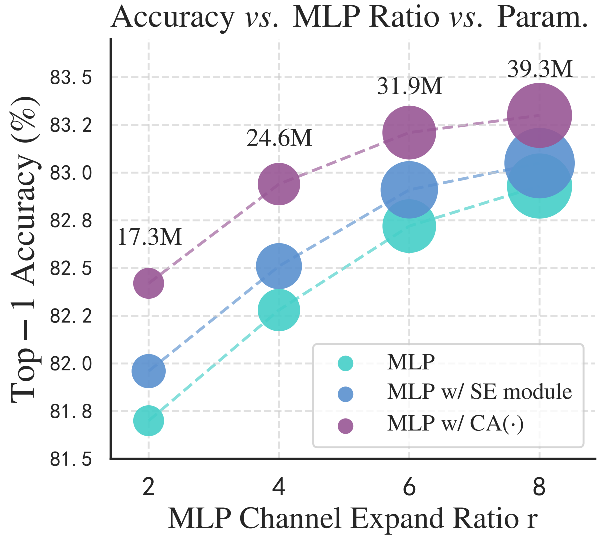
$$Y = \mathrm{GELU}\Big(\mathrm{DW}_{3\times3}\big(\mathrm{Conv}_{1\times1}(\mathrm{Norm}(X))\big)\Big),$$

$$Z = \mathrm{Conv}_{1\times1}(\mathrm{CA}(Y)) + X.$$

$$\mathrm{CA}(X) = X + \gamma_c \odot \big(X - \mathrm{GELU}(XW_r)\big)$$



Channel energy ranks and channel saliency maps (CSM)[1]



Modules		Top-1 Acc (%)	Params. (M)	FLOPs (G)
Baseline		76.6	4.75	1.01
SMixer	+Gating branch	77.3	5.09	1.07
	+$\mathrm{DW}_{7\times7}$	77.5	5.14	1.09
	+Multi-order DW(·)	78.0	5.17	1.10
	+FD(·)	78.3	5.18	1.10
CMixer	+SE module	78.6	5.29	1.14
	+CA(·)	**79.0**	5.20	1.10

Ablation of MogaNet-S on ImageNet

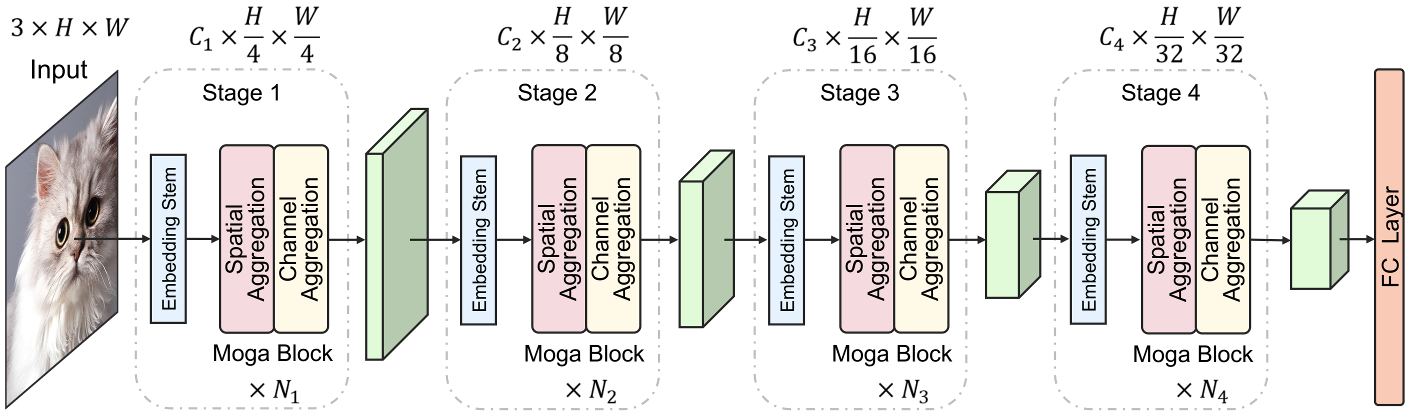
[1] Reflash dropout in image supe-resolution. CVPR, 2022.

The 12th Internaltional Conference on Learning Representations

MogaNet: Configurations

4-stage network designs.

- 3-5M: X-Tiny & Tiny
- 25-50M: Small & Base
- 80-200M: Large & X-Large



Stage	Output Size	Layer Settings	MogaNet XTiny	Tiny	Small	Base	Large	XLarge
S1	$\frac{H \times W}{4 \times 4}$	Stem	$\text{Conv}_{3\times3}$, stride 2, C/2 $\text{Conv}_{3\times3}$, stride 2, C					
		Embed. Dim.	32	32	64	64	64	96
		# Moga Block	3	3	2	4	4	6
		MLP Ratio	8					
S2	$\frac{H \times W}{8 \times 8}$	Stem	$\text{Conv}_{3\times3}$, stride 2					
		Embed. Dim.	64	64	128	160	160	192
		# Moga Block	3	3	3	6	6	6
		MLP Ratio	8					
S3	$\frac{H \times W}{16 \times 16}$	Stem	$\text{Conv}_{3\times3}$, stride 2					
		Embed. Dim.	96	128	320	320	320	480
		# Moga Block	10	12	12	22	44	44
		MLP Ratio	4					
S4	$\frac{H \times W}{32 \times 32}$	Stem	$\text{Conv}_{3\times3}$, stride 2					
		Embed. Dim.	192	256	512	512	640	960
		# Moga Block	2	2	2	3	4	4
		MLP Ratio	4					
Classifier			Global Average Pooling, Linear					
Parameters (M)			2.97	5.20	25.3	43.8	82.5	180.8
FLOPs (G)			0.80	1.10	4.97	9.93	15.9	34.5

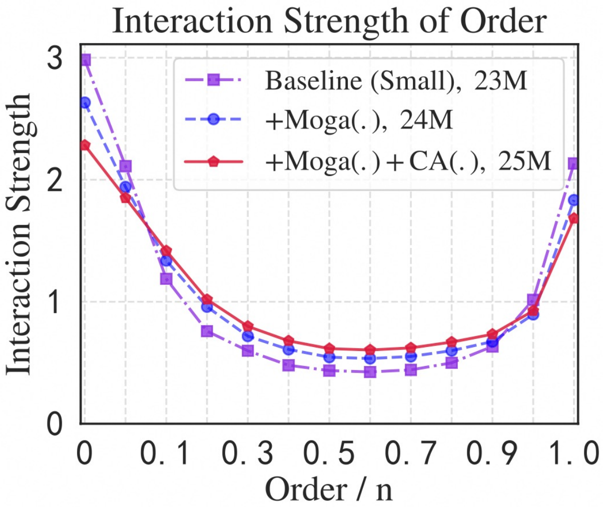
Configuration	DeiT	RSB A2	MogaNet XT	T	S	B	L	XL
Input resolution	224^2	224^2	224^2					
Epochs	300	300	300					
Batch size	1024	2048	1024					
Optimizer	AdamW	LAMB	AdamW					
AdamW (β_1, β_2)	$0.9, 0.999$	-	$0.9, 0.999$					
Learning rate	0.001	0.005	0.001					
Learning rate decay	Cosine	Cosine	Cosine					
Weight decay	0.05	0.02	0.03	0.04	0.05	0.05	0.05	0.05
Warmup epochs	5	5	5					
Label smoothing ϵ	0.1	0.1	0.1					
Stochastic Depth	✓	✓	0.05	0.1	0.1	0.2	0.3	0.4
Rand Augment	9/0.5	7/0.5	7/0.5	7/0.5	9/0.5	9/0.5	9/0.5	9/0.5
Repeated Augment	✓	✓	✗					
Mixup α	0.8	0.1	0.1	0.1	0.8	0.8	0.8	0.8
CutMix α	1.0	1.0	1.0					
Erasing prob.	0.25	✗	0.25					
ColorJitter	✗	✗	✗	✗	0.4	0.4	0.4	0.4
Gradient Clipping	✓	✗	✗					
EMA decay	✓	✗	✗	✗	✓	✓	✓	✓
Test crop ratio	0.875	0.95	0.90					

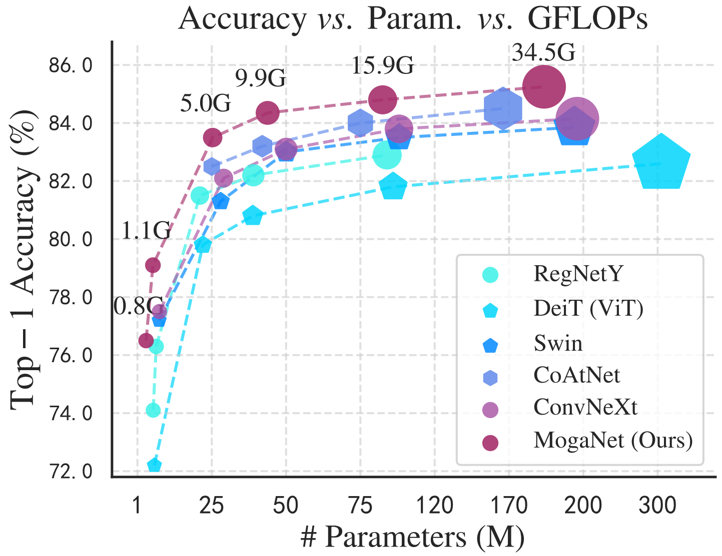
Configuration	IN-21K PT S	B	L	XL	IN-1K FT S	B	L	XL
Input resolution	224^2				384^2			
Epochs	90				30			
Batch size	1024				512			
Optimizer	AdamW				AdamW			
AdamW (β_1, β_2)	$0.9, 0.999$				$0.9, 0.999$			
Learning rate	1×10^{-3}				5×10^{-5}			
Learning rate decay	Cosine				Cosine			
Weight decay	0.05				0.05			
Warmup epochs	5				0			
Label smoothing ϵ	0.2				0.1	0.1	0.2	0.2
Stochastic Depth	0	0.1	0.1	0.1	0.4	0.6	0.7	0.8
Rand Augment	9/0.5				9/0.5			
Repeated Augment	✗				✗			
Mixup α	0.8				✗			
CutMix α	1.0				✗			
Erasing prob.	0.25				0.25			
ColorJitter	0.4				0.4			
Gradient Clipping	✗				✗			
EMA decay	✗				✓			
Test crop ratio	0.90				1.0			

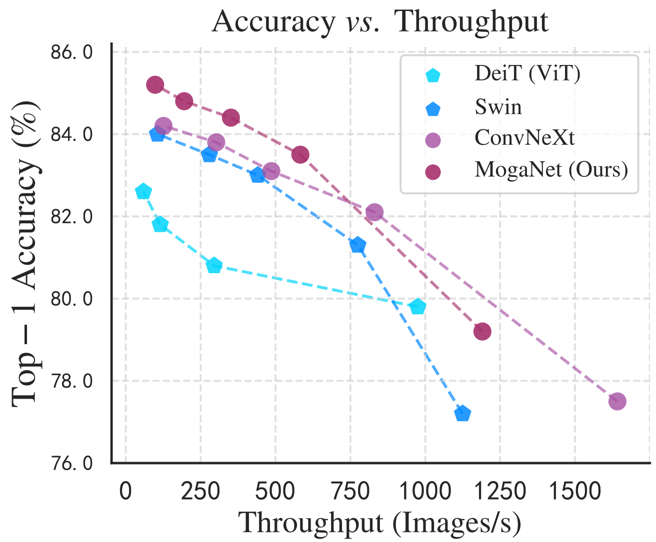
MogaNet: Ablation and Analysis

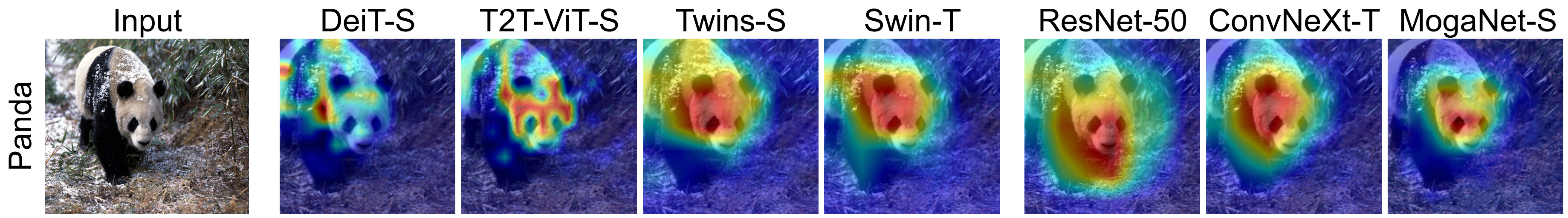
- Great scalability and efficiency of parameter usage.
- Relieving representation bottleneck.

Modules	Top-1 Acc (%)
ConvNeXt-T	82.1
Baseline	82.2
Moga Block	**83.4**
−FD(·)	83.2
−Multi-DW(·)	83.1
−Moga(·)	82.7
−CA(·)	82.9









Grad-CAM visualization

MogaNet: ImageNet Classification

Light weight (3-10M)

Architecture	Date	Type	Image Size	Param. (M)	FLOPs (G)	Top-1 Acc (%)
ResNet-18	CVPR'2016	C	224^2	11.7	1.80	71.5
ShuffleNetV2 2×	ECCV'2018	C	224^2	5.5	0.60	75.4
EfficientNet-B0	ICML'2019	C	224^2	5.3	0.39	77.1
RegNetY-800MF	CVPR'2020	C	224^2	6.3	0.80	76.3
DeiT-T†	ICML'2021	T	224^2	5.7	1.08	74.1
PVT-T	ICCV'2021	T	224^2	13.2	1.60	75.1
T2T-ViT-7	ICCV'2021	T	224^2	4.3	1.20	71.7
ViT-C	NIPS'2021	T	224^2	4.6	1.10	75.3
SReT-$T_{Distill}$	ECCV'2022	T	224^2	4.8	1.10	77.6
PiT-Ti	ICCV'2021	H	224^2	4.9	0.70	74.6
LeViT-S	ICCV'2021	H	224^2	7.8	0.31	76.6
CoaT-Lite-T	ICCV'2021	H	224^2	5.7	1.60	77.5
Swin-1G	ICCV'2021	H	224^2	7.3	1.00	77.3
MobileViT-S	ICLR'2022	H	256^2	5.6	4.02	78.4
MobileFormer-294M	CVPR'2022	H	224^2	11.4	0.59	77.9
ConvNext-XT	CVPR'2022	C	224^2	7.4	0.60	77.5
VAN-B0	CVMJ'2023	C	224^2	4.1	0.88	75.4
ParC-Net-S	ECCV'2022	C	256^2	5.0	3.48	78.6
MogaNet-XT	Ours	C	256^2	3.0	1.04	77.2
MogaNet-T	Ours	C	224^2	5.2	1.10	79.0
MogaNet-T§	Ours	C	256^2	5.2	1.44	**80.0**

Architecture	Input size	Learning rate	Warmup epochs	Rand Augment	3-Augment	EMA	Top-1 Acc (%)
MogaNet-XT	224^2	1×10^{-3}	5	7/0.5	✗	✗	76.5
MogaNet-XT	224^2	2×10^{-3}	20	✗	✓	✗	77.1
MogaNet-XT	256^2	1×10^{-3}	5	7/0.5	✗	✗	77.2
MogaNet-XT	256^2	2×10^{-3}	20	✗	✓	✗	77.6
MogaNet-T	224^2	1×10^{-3}	5	7/0.5	✗	✗	79.0
MogaNet-T	224^2	2×10^{-3}	20	✗	✓	✗	79.4
MogaNet-T	256^2	1×10^{-3}	5	7/0.5	✗	✗	79.6
MogaNet-T	256^2	2×10^{-3}	20	✗	✓	✗	**80.0**

Normal size (25-50M)

Architecture	Date	Type	Image Size	Param. (M)	FLOPs (G)	Top-1 Acc (%)
Deit-S	ICML'2021	T	224^2	22	4.6	79.8
Swin-T	ICCV'2021	T	224^2	28	4.5	81.3
CSWin-T	CVPR'2022	T	224^2	23	4.3	82.8
LITV2-S	NIPS'2022	T	224^2	28	3.7	82.0
CoaT-S	ICCV'2021	H	224^2	22	12.6	82.1
CoAtNet-0	NIPS'2021	H	224^2	25	4.2	82.7
UniFormer-S	ICLR'2022	H	224^2	22	3.6	82.9
RegNetY-4GF†	CVPR'2020	C	224^2	21	4.0	81.5
ConvNeXt-T	CVPR'2022	C	224^2	29	4.5	82.1
SLaK-T	ICLR'2023	C	224^2	30	5.0	82.5
HorNet-$T_{7\times7}$	NIPS'2022	C	224^2	22	4.0	82.8
MogaNet-S	Ours	C	224^2	25	5.0	**83.4**
Swin-S	ICCV'2021	T	224^2	50	8.7	83.0
Focal-S	NIPS'2021	T	224^2	51	9.1	83.6
CSWin-S	CVPR'2022	T	224^2	35	6.9	83.6
LITV2-M	NIPS'2022	T	224^2	49	7.5	83.3
CoaT-M	ICCV'2021	H	224^2	45	9.8	83.6
CoAtNet-1	NIPS'2021	H	224^2	42	8.4	83.3
UniFormer-B	ICLR'2022	H	224^2	50	8.3	83.9
FAN-B-Hybrid	ICML'2022	H	224^2	50	11.3	83.9
EfficientNet-B6	ICML'2019	C	528^2	43	19.0	84.0
RegNetY-8GF†	CVPR'2020	C	224^2	39	8.1	82.2
ConvNeXt-S	CVPR'2022	C	224^2	50	8.7	83.1
FocalNet-S (LRF)	NIPS'2022	C	224^2	50	8.7	83.5
HorNet-$S_{7\times7}$	NIPS'2022	C	224^2	50	8.8	84.0
SLaK-S	ICLR'2023	C	224^2	55	9.8	83.8
MogaNet-B	Ours	C	224^2	44	9.9	**84.3**

Training and inference at the resolution of 224^2 or 256^2.

Large size (80-200M)

DeiT-B	ICML'2021	T	224^2	86	17.5	81.8
Swin-B	ICCV'2021	T	224^2	89	15.4	83.5
Focal-B	NIPS'2021	T	224^2	90	16.4	84.0
CSWin-B	CVPR'2022	T	224^2	78	15.0	84.2
DeiT III-B	ECCV'2022	T	224^2	87	18.0	83.8
BoTNet-T7	CVPR'2021	H	256^2	79	19.3	84.2
CoAtNet-2	NIPS'2021	H	224^2	75	15.7	84.1
FAN-B-Hybrid	ICML'2022	H	224^2	77	16.9	84.3
RegNetY-16GF	CVPR'2020	C	224^2	84	16.0	82.9
ConvNeXt-B	CVPR'2022	C	224^2	89	15.4	83.8
RepLKNet-31B	CVPR'2022	C	224^2	79	15.3	83.5
FocalNet-B (LRF)	NIPS'2022	C	224^2	89	15.4	83.9
HorNet-$B_{7\times7}$	NIPS'2022	C	224^2	87	15.6	84.3
SLaK-B	ICLR'2023	C	224^2	95	17.1	84.0
MogaNet-L	Ours	C	224^2	83	15.9	**84.7**
Swin-L‡	ICCV'2021	T	384^2	197	104	87.3
DeiT III-L‡	ECCV'2022	T	384^2	304	191	87.7
CoAtNet-3‡	NIPS'2021	H	384^2	168	107	87.6
RepLKNet-31L‡	CVPR'2022	C	384^2	172	96	86.6
ConvNeXt-L	CVPR'2022	C	224^2	198	34.4	84.3
ConvNeXt-L‡	CVPR'2022	C	384^2	198	101	87.5
ConvNeXt-XL‡	CVPR'2022	C	384^2	350	179	87.8
HorNet-L‡	NIPS'2022	C	384^2	202	102	87.7
MogaNet-XL	Ours	C	224^2	181	34.5	85.1
MogaNet-XL‡	Ours	C	384^2	181	102	**87.8**

Architecture	Date	Type	Param. (M)	100-epoch Train	100-epoch Test	100-epoch Acc (%)	300-epoch Train	300-epoch Test	300-epoch Acc (%)
ConvNeXt-T (Liu et al., 2022b)	CVPR'2022	C	29	160^2	224^2	78.8	224^2	224^2	82.1
ConvNeXt-S (Liu et al., 2022b)	CVPR'2022	C	50	160^2	224^2	81.7	224^2	224^2	83.1
ConvNeXt-B (Liu et al., 2022b)	CVPR'2022	C	89	160^2	224^2	82.1	224^2	224^2	83.8
ConvNeXt-L (Liu et al., 2022b)	CVPR'2022	C	189	160^2	224^2	82.8	224^2	224^2	84.3
ConvNeXt-XL (Liu et al., 2022b)	CVPR'2022	C	350	160^2	224^2	82.9	224^2	224^2	84.5
HorNet-$T_{7\times7}$ (Rao et al., 2022)	NIPS'2022	C	22	160^2	224^2	80.1	224^2	224^2	82.8
HorNet-$S_{7\times7}$ (Rao et al., 2022)	NIPS'2022	C	50	160^2	224^2	81.2	224^2	224^2	84.0
VAN-B0 (Guo et al., 2023)	CVMJ'2023	C	4	160^2	224^2	72.6	224^2	224^2	75.8
VAN-B2 (Guo et al., 2023)	CVMJ'2023	C	27	160^2	224^2	81.0	224^2	224^2	82.8
VAN-B3 (Guo et al., 2023)	CVMJ'2023	C	45	160^2	224^2	81.9	224^2	224^2	83.9
MogaNet-XT	Ours	C	3	160^2	224^2	72.8	224^2	224^2	76.5
MogaNet-T	Ours	C	5	160^2	224^2	75.4	224^2	224^2	79.0
MogaNet-S	Ours	C	25	160^2	224^2	81.1	224^2	224^2	83.4
MogaNet-B	Ours	C	44	160^2	224^2	82.2	224^2	224^2	84.3
MogaNet-L	Ours	C	83	160^2	224^2	83.2	224^2	224^2	84.7

MogaNet: COCO Object Det. and Seg.

RetinaNet (1×)

Architecture	Type	#P. (M)	FLOPs (G)	RetinaNet 1× AP	AP_{50}	AP_{75}	AP^S	AP_M	AP_L
RegNet-800M	C	17	168	35.6	54.7	37.7	19.7	390	47.8
PVTV2-B0	T	13	160	37.1	57.2	39.2	23.4	40.4	49.2
MogaNet-XT	C	12	167	**39.7**	**60.0**	**42.4**	**23.8**	**43.6**	**51.7**
ResNet-18	C	21	189	31.8	49.6	33.6	16.3	34.3	43.2
RegNet-1.6G	C	20	185	37.4	56.8	39.8	22.4	41.1	49.2
RegNet-3.2G	C	26	218	39.0	58.4	41.9	22.6	43.5	50.8
PVT-T	T	23	183	36.7	56.9	38.9	22.6	38.8	50.0
PoolFormer-S12	T	22	207	36.2	56.2	38.2	20.8	39.1	48.0
PVTV2-B1	T	24	187	41.1	61.4	43.8	26.0	44.6	54.6
MogaNet-T	C	14	173	**41.4**	**61.5**	**44.4**	**25.1**	**45.7**	**53.6**
ResNet-50	C	37	239	36.3	55.3	38.6	19.3	40.0	48.8
Swin-T	T	38	245	41.8	62.6	44.7	25.2	45.8	54.7
PVT-S	T	34	226	40.4	61.3	43.0	25.0	42.9	55.7
Twins-SVT-S	T	34	209	42.3	63.4	45.2	26.0	45.5	56.5
Focal-T	T	39	265	43.7	-	-	-	-	-
PoolFormer-S36	T	41	272	39.5	60.5	41.8	22.5	42.9	52.4
PVTV2-B2	T	35	281	44.6	65.7	47.6	28.6	48.5	59.2
CMT-S	H	45	231	44.3	65.5	47.5	27.1	48.3	59.1
MogaNet-S	C	35	253	**45.8**	**66.6**	**49.0**	**29.1**	**50.1**	**59.8**
ResNet-101	C	57	315	38.5	57.8	41.2	21.4	42.6	51.1
PVT-M	T	54	258	41.9	63.1	44.3	25.0	44.9	57.6
Focal-S	T	62	367	45.6	-	-	-	-	-
PVTV2-B3	T	55	263	46.0	67.0	49.5	28.2	50.0	61.3
PVTV2-B4	T	73	315	46.3	67.0	49.6	29.0	50.1	62.7
MogaNet-B	C	54	355	**47.7**	**68.9**	**51.0**	**30.5**	**52.2**	**61.7**
ResNeXt-101-64	C	95	473	41.0	60.9	44.0	23.9	45.2	54.0
PVTV2-B5	T	92	335	46.1	66.6	49.5	27.8	50.2	62.0
MogaNet-L	C	92	477	**48.7**	**69.5**	**52.6**	**31.5**	**53.4**	**62.7**

Inference input size 800×1280

Mask R-CNN (1×)

Architecture	Type	#P. (M)	FLOPs (G)	Mask R-CNN 1× AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
RegNet-800M	C	27	187	37.5	57.9	41.1	34.3	56.0	36.8
MogaNet-XT	C	23	185	**40.7**	**62.3**	**44.4**	**37.6**	**59.6**	**40.2**
ResNet-18	C	31	207	34.0	54.0	36.7	31.2	51.0	32.7
RegNet-1.6G	C	29	204	38.9	60.5	43.1	35.7	57.4	38.9
PVT-T	T	33	208	36.7	59.2	39.3	35.1	56.7	37.3
PoolFormer-S12	T	32	207	37.3	59.0	40.1	34.6	55.8	36.9
MogaNet-T	C	25	192	**42.6**	**64.0**	**46.4**	**39.1**	**61.3**	**42.0**
ResNet-50	C	44	260	38.0	58.6	41.4	34.4	55.1	36.7
RegNet-6.4G	C	45	307	41.1	62.3	45.2	37.1	59.2	39.6
PVT-S	T	44	245	40.4	62.9	43.8	37.8	60.1	40.3
Swin-T	T	48	264	42.2	64.6	46.2	39.1	61.6	42.0
MViT-T	T	46	326	45.9	**68.7**	50.5	42.1	**66.0**	45.4
PoolFormer-S36	T	32	207	41.0	63.1	44.8	37.7	60.1	40.0
Focal-T	T	49	291	44.8	67.7	49.2	41.0	64.7	44.2
PVTV2-B2	T	45	309	45.3	67.1	49.6	41.2	64.2	44.4
LITV2-S	T	47	261	44.9	67.0	49.5	40.8	63.8	44.2
CMT-S	H	45	249	44.6	66.8	48.9	40.7	63.9	43.4
Conformer-S/16	H	58	341	43.6	65.6	47.7	39.7	62.6	42.5
Uniformer-S	H	41	269	45.6	68.1	49.7	41.6	64.8	45.0
ConvNeXt-T	C	48	262	44.2	66.6	48.3	40.1	63.3	42.8
FocalNet-T (SRF)	C	49	267	45.9	68.3	50.1	41.3	65.0	44.3
FocalNet-T (LRF)	C	49	268	46.1	68.2	50.6	41.5	65.1	44.5
MogaNet-S	C	45	272	**46.7**	68.0	**51.3**	**42.2**	65.4	**45.5**
ResNet-101	C	63	336	40.4	61.1	44.2	36.4	57.7	38.8
RegNet-12G	C	64	423	42.2	63.7	46.1	38.0	60.5	40.5
PVT-M	T	64	302	42.0	64.4	45.6	39.0	61.6	42.1
Swin-S	T	69	354	44.8	66.6	48.9	40.9	63.4	44.2
Focal-S	T	71	401	47.4	69.8	51.9	42.8	66.6	46.1
PVTV2-B3	T	65	397	47.0	68.1	51.7	42.5	65.7	45.7
LITV2-M	T	68	315	46.5	68.0	50.9	42.0	65.1	45.0
UniFormer-B	H	69	399	47.4	69.7	52.1	43.1	66.0	46.5
ConvNeXt-S	C	70	348	45.4	67.9	50.0	41.8	65.2	45.1
MogaNet-B	C	63	373	**47.9**	**70.0**	**52.7**	**43.2**	**67.0**	**46.6**
Swin-B	T	107	496	46.9	69.6	51.2	42.3	65.9	45.6
PVTV2-B5	T	102	557	47.4	68.6	51.9	42.5	65.7	46.0
ConvNeXt-B	C	108	486	47.0	69.4	51.7	42.7	66.3	46.0
FocalNet-B (SRF)	C	109	496	48.8	70.7	53.5	43.3	67.5	46.5
MogaNet-L	C	102	495	**49.4**	**70.7**	**54.1**	**44.1**	**68.1**	**47.6**

Cascade Mask R-CNN (3×)

Architecture	Type	#P. (M)	FLOPs (G)	Cascade Mask R-CNN +MS 3× AP^{bb}	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet-50	C	77	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T	T	86	745	50.4	69.2	54.7	43.7	66.6	47.3
Focal-T	T	87	770	51.5	70.6	55.9	-	-	-
ConvNeXt-T	C	86	741	50.4	69.1	54.8	43.7	66.5	47.3
FocalNet-T (SRF)	C	86	746	51.5	70.1	55.8	44.6	67.7	48.4
MogaNet-S	C	78	750	**51.6**	**70.8**	**56.3**	**45.1**	**68.7**	**48.8**
ResNet-101-32	C	96	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S	T	107	838	51.9	70.7	56.3	45.0	68.2	48.8
ConvNeXt-S	C	108	827	51.9	70.8	56.5	45.0	68.4	49.1
MogaNet-B	C	101	851	**52.6**	**72.0**	**57.3**	**46.0**	**69.6**	**49.7**
Swin-B	T	145	982	51.9	70.5	56.4	45.0	68.1	48.9
ConvNeXt-B	C	146	964	52.7	71.3	57.2	45.6	68.9	49.5
MogaNet-L	C	140	974	**53.3**	**71.8**	**57.8**	**46.1**	**69.2**	**49.8**
Swin-L‡	T	253	1382	53.9	72.4	58.8	46.7	70.1	50.8
ConvNeXt-L‡	C	255	1354	54.8	73.8	59.8	47.6	71.3	51.7
ConvNeXt-XL‡	C	407	1898	55.2	74.2	59.9	47.7	71.6	52.2
RepLKNet-31L‡	C	229	1321	53.9	72.5	58.6	46.5	70.0	50.6
HorNet-L‡	C	259	1399	56.0	-	-	48.6	-	-
MogaNet-XL‡	C	238	1355	**56.2**	**75.0**	**61.2**	**48.8**	**72.6**	**53.3**

- Object Detection: RetinaNet.
- Instance Segmentation: (Cascade) Mask R-CNN.
- Multi-scale fine-tuning with IN-21K pre-trained models.
- Codebase: MMDetection.

MogaNet: ADE20K Semantic Segmentation

The 12th Internaltional Conference on Learning Representations

ADE20K Semantic FPN (80K)

Method	Architecture	Date	Crop size	Param. (M)	FLOPs (G)	mIoU^{ss} (%)
Semantic FPN (80K)	PVT-S	ICCV'2021	512^2	28	161	39.8
	Twins-S	NIPS'2021	512^2	28	162	44.3
	Swin-T	ICCV'2021	512^2	32	182	41.5
	Uniformer-S	ICLR'2022	512^2	25	247	46.6
	LITV2-S	NIPS'2022	512^2	31	179	44.3
	VAN-B2	CVMJ'2023	512^2	30	164	46.7
	MogaNet-S	Ours	512^2	29	189	**47.7**

MogaNet + Semantic FPN

Method	Backbone	Pretrain	Params	FLOPs	Iters	mIoU	mAcc
Semantic FPN	MogaNet-XT	ImageNet-1K	6.9M	101.4G	80K	40.3	52.4
Semantic FPN	MogaNet-T	ImageNet-1K	9.1M	107.8G	80K	43.1	55.4
Semantic FPN	MogaNet-S	ImageNet-1K	29.1M	189.7G	80K	47.7	59.8
Semantic FPN	MogaNet-B	ImageNet-1K	47.5M	293.6G	80K	49.3	61.6
Semantic FPN	MogaNet-L	ImageNet-1K	86.2M	418.7G	80K	50.2	63.0

- Semantic FPN (80K) with 512×2048 inference sizes.
- UperNet (160K) with 512×2048 or 640×2560 inference resolutions using IN-1K or IN-21K models.
- Codebase: MMSegmentation.

ADE20K UperNet (160K)

Architecture	Date	Type	Crop size	Param. (M)	FLOPs (G)	mIoU^{ss} (%)
ResNet-18	CVPR'2016	C	512^2	41	885	39.2
MogaNet-XT	Ours	C	512^2	30	856	**42.2**
ResNet-50	CVPR'2016	C	512^2	67	952	42.1
MogaNet-T	Ours	C	512^2	33	862	**43.7**
DeiT-S	ICML'2021	T	512^2	52	1099	44.0
Swin-T	ICCV'2021	T	512^2	60	945	46.1
TwinsP-S	NIPS'2021	T	512^2	55	919	46.2
Twins-S	NIPS'2021	T	512^2	54	901	46.2
Focal-T	NIPS'2021	T	512^2	62	998	45.8
Uniformer-S$_{h32}$	ICLR'2022	H	512^2	52	955	47.0
UniFormer-S	ICLR'2022	H	512^2	52	1008	47.6
ConvNeXt-T	CVPR'2022	C	512^2	60	939	46.7
FocalNet-T (SRF)	NIPS'2022	C	512^2	61	944	46.5
HorNet-T$_{7\times7}$	NIPS'2022	C	512^2	52	926	48.1
MogaNet-S	Ours	C	512^2	55	946	**49.2**
Swin-S	ICCV'2021	T	512^2	81	1038	48.1
Twins-B	NIPS'2021	T	512^2	89	1020	47.7
Focal-S	NIPS'2021	T	512^2	85	1130	48.0
Uniformer-B$_{h32}$	ICLR'2022	H	512^2	80	1106	49.5
ConvNeXt-S	CVPR'2022	C	512^2	82	1027	48.7
FocalNet-S (SRF)	NIPS'2022	C	512^2	83	1035	49.3
SLaK-S	ICLR'2023	C	512^2	91	1028	49.4
MogaNet-B	Ours	C	512^2	74	1050	**50.1**
Swin-B	ICCV'2021	T	512^2	121	1188	49.7
Focal-B	NIPS'2021	T	512^2	126	1354	49.0
ConvNeXt-B	CVPR'2022	C	512^2	122	1170	49.1
RepLKNet-31B	CVPR'2022	C	512^2	112	1170	49.9
FocalNet-B (SRF)	NIPS'2022	C	512^2	124	1180	50.2
SLaK-B	ICLR'2023	C	512^2	135	1185	50.2
MogaNet-L	Ours	C	512^2	113	1176	**50.9**
Swin-L‡	ICCV'2021	T	640^2	234	2468	52.1
ConvNeXt-L‡	CVPR'2022	C	640^2	245	2458	53.7
RepLKNet-31L‡	CVPR'2022	C	640^2	207	2404	52.4
MogaNet-XL‡	Ours	C	640^2	214	2451	**54.0**

MogaNet: 2D/3D Human Pose Estimation

ICLR
The 12th Internaltional Conference on Learning Representations

COCO 2D Human Pose with TopDown baseline (256×192)

Architecture	Type	Crop size	#P. (M)	FLOPs (G)	AP (%)	AP^{50} (%)	AP^{75} (%)	AR (%)
MobileNetV2	C	256 × 192	10	1.6	64.6	87.4	72.3	70.7
ShuffleNetV2 2×	C	256 × 192	8	1.4	59.9	85.4	66.3	66.4
MogaNet-XT	C	256 × 192	6	1.8	**72.1**	**89.7**	**80.1**	**77.7**
RSN-18	C	256 × 192	9	2.3	70.4	88.7	77.9	77.1
MogaNet-T	C	256 × 192	8	2.2	**73.2**	**90.1**	**81.0**	**78.8**
ResNet-50	C	256 × 192	34	5.5	72.1	89.9	80.2	77.6
HRNet-W32	C	256 × 192	29	7.1	74.4	90.5	81.9	78.9
Swin-T	T	256 × 192	33	6.1	72.4	90.1	80.6	78.2
PVT-S	T	256 × 192	28	4.1	71.4	89.6	79.4	77.3
PVTV2-B2	T	256 × 192	29	4.3	73.7	90.5	81.2	79.1
Uniformer-S	H	256 × 192	25	4.7	74.0	90.3	82.2	79.5
ConvNeXt-T	C	256 × 192	33	5.5	73.2	90.0	80.9	78.8
MogaNet-S	C	256 × 192	29	6.0	**74.9**	**90.7**	**82.8**	**80.1**
ResNet-101	C	256 × 192	53	12.4	71.4	89.3	79.3	77.1
ResNet-152	C	256 × 192	69	15.7	72.0	89.3	79.8	77.8
HRNet-W48	C	256 × 192	64	14.6	75.1	90.6	82.2	80.4
Swin-B	T	256 × 192	93	18.6	72.9	89.9	80.8	78.6
Swin-L	T	256 × 192	203	40.3	74.3	90.6	82.1	79.8
Uniformer-B	H	256 × 192	54	9.2	75.0	90.6	83.0	80.4
ConvNeXt-S	C	256 × 192	55	9.7	73.7	90.3	81.9	79.3
ConvNeXt-B	C	256 × 192	94	16.4	74.0	90.7	82.1	79.5
MogaNet-B	C	256 × 192	47	10.9	**75.3**	**90.9**	**83.3**	**80.7**

COCO 2D Human Pose with TopDown baseline (384×288)

Architecture	Type	Crop size	#P. (M)	FLOPs (G)	AP (%)	AP^{50} (%)	AP^{75} (%)	AR (%)
MobileNetV2	C	384 × 288	10	3.6	67.3	87.9	74.3	72.9
ShuffleNetV2 2×	C	384 × 288	8	3.1	63.6	86.5	70.5	69.7
MogaNet-XT	C	384 × 288	6	4.2	**74.7**	**90.1**	**81.3**	**79.9**
RSN-18	C	384 × 288	9	5.1	72.1	89.5	79.8	78.6
MogaNet-T	C	384 × 288	8	4.9	**75.7**	**90.6**	**82.6**	**80.9**
HRNet-W32	C	384 × 288	29	16.0	75.8	90.6	82.7	81.0
Uniformer-S	H	384 × 288	25	11.1	75.9	90.6	83.4	81.4
ConvNeXt-T	C	384 × 288	33	33.1	75.3	90.4	82.1	80.5
MogaNet-S	C	384 × 288	29	13.5	**76.4**	**91.0**	**83.3**	**81.4**
ResNet-152	C	384 × 288	69	35.6	74.3	89.6	81.1	79.7
HRNet-W48	C	384 × 288	64	32.9	76.3	90.8	82.0	81.2
Swin-B	T	384 × 288	93	39.2	74.9	90.5	81.8	80.3
Swin-L	T	384 × 288	203	86.9	76.3	91.2	83.0	814
HRFormer-B	T	384 × 288	54	30.7	77.2	91.0	83.6	82.0
ConvNeXt-S	C	384 × 288	55	21.8	75.8	90.7	83.1	81.0
ConvNeXt-B	C	384 × 288	94	36.6	75.9	90.6	83.1	81.1
Uniformer-B	C	384 × 288	54	14.8	76.7	90.8	84.0	81.4
MogaNet-B	C	384 × 288	47	24.4	**77.3**	**91.4**	**84.0**	**82.2**

3D Human Pose with Expose

- 3D Face: FFHQ (256^2)
- 3D Hand: FreiHand (224^2)
- Codebase: MMPose

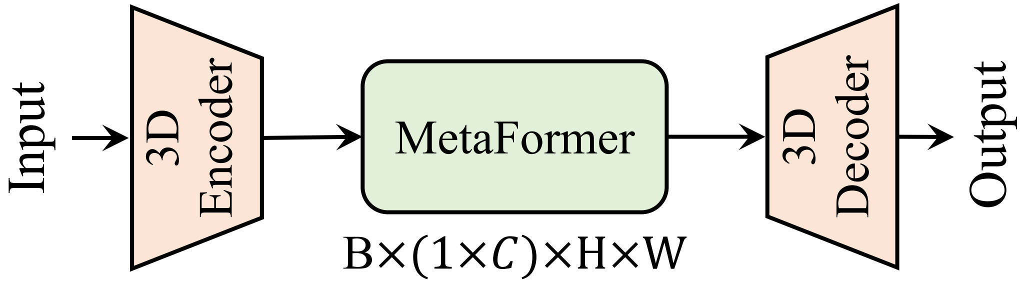
Architecture	Type	Hand #P. (M)	Hand FLOPs (G)	Hand PA-MPJPE (mm)↓	Face #P. (M)	Face FLOPs (G)	Face 3DRMSE ↓
MobileNetV2	C	4.8	0.3	8.33	4.9	0.4	2.64
ResNet-18	C	13.0	1.8	7.51	13.1	2.4	2.40
MogaNet-T	C	6.5	1.1	**6.82**	6.6	1.5	**2.36**
ResNet-50	C	26.9	4.1	6.85	27.0	5.4	2.48
ResNet-101	C	45.9	7.9	6.44	46.0	10.3	2.47
DeiT-S	T	23.4	4.3	7.86	23.5	5.5	2.52
Swin-T	T	30.2	4.6	6.97	30.3	6.1	2.45
Swin-S	T	51.0	13.8	6.50	50.9	8.5	2.48
ConvNeXt-T	C	29.9	4.5	6.18	30.0	5.8	2.34
ConvNeXt-S	C	51.5	8.7	6.04	51.6	11.4	2.27
HorNet-T	C	23.7	4.3	6.46	23.8	5.6	2.39
MogaNet-S	C	26.6	5.0	**6.08**	26.7	6.5	**2.24**

MogaNet: Video Prediction

Moving MNIST (10×1×64×64)

Architecture	#P. (M)	FLOPs (G)	FPS (s)	200 epochs MSE↓	200 epochs MAE↓	200 epochs SSIM↑	2000 epochs MSE↓	2000 epochs MAE↓	2000 epochs SSIM↑
ViT	46.1	16.9	290	35.15	95.87	0.9139	19.74	61.65	0.9539
Swin	46.1	16.4	294	29.70	84.05	0.9331	19.11	59.84	0.9584
Uniformer	44.8	16.5	296	30.38	85.87	0.9308	18.01	57.52	0.9609
MLP-Mixer	38.2	14.7	334	29.52	83.36	0.9338	18.85	59.86	0.9589
ConvMixer	3.9	5.5	658	32.09	88.93	0.9259	22.30	67.37	0.9507
Poolformer	37.1	14.1	341	31.79	88.48	0.9271	20.96	64.31	0.9539
SimVP	58.0	19.4	209	32.15	89.05	0.9268	21.15	64.15	0.9536
ConvNeXt	37.3	14.1	344	26.94	77.23	0.9397	17.58	55.76	0.9617
VAN	44.5	16.0	288	26.10	76.11	0.9417	16.21	53.57	0.9646
HorNet	45.7	16.3	287	29.64	83.26	0.9331	17.40	55.70	0.9624
MogaNet	46.8	16.5	255	**25.57**	**75.19**	**0.9429**	**15.67**	**51.84**	**0.9661**

- Replacing the MetaFormer blocks in SimVP.
- Comparison with MMNIST and MMNIST-CIFAR.



Moving MNIST-CIFAR (10×3×64×64)

	Method	Params (M)	FLOPs (G)	FPS	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑
Recurrent-based	ConvLSTM	15.0	56.8	113	73.31	338.56	0.9204	23.09
	PredNet	12.5	8.4	659	286.70	514.14	0.8139	17.49
	PredRNN	23.8	116.0	54	50.09	225.04	0.9499	24.90
	PredRNN++	38.6	171.7	38	**44.19**	198.27	**0.9567**	**25.60**
	MIM	38.0	179.2	37	48.63	213.44	0.9521	25.08
	E3D-LSTM	51.0	298.9	18	80.79	214.86	0.9314	22.89
	PhyDNet	3.1	15.3	182	142.54	700.37	0.8276	19.92
	MAU	4.5	17.8	201	58.84	255.76	0.9408	24.19
	PredRNNv2	23.9	116.6	52	57.27	252.29	0.9419	24.24
	DMVFN	3.5	0.2	1145	298.73	606.92	0.7765	17.07
Recurrent-free	SimVP	58.0	19.4	209	59.83	214.54	0.9414	24.15
	TAU	44.7	16.0	283	48.17	**177.35**	0.9539	25.21
	SimVPv2	46.8	16.5	282	51.13	185.13	0.9512	24.93
	ViT	46.1	16.9	290	64.94	234.01	0.9354	23.90
	Swin Transformer	46.1	16.4	294	57.11	207.45	0.9443	24.34
	Uniformer	44.8	16.5	296	56.96	207.51	0.9442	24.38
	MLP-Mixer	38.2	14.7	334	57.03	206.46	0.9446	24.34
	ConvMixer	3.9	5.5	658	59.29	219.76	0.9403	24.17
	Poolformer	37.1	14.1	341	60.98	219.50	0.9399	24.16
	ConvNext	37.3	14.1	344	51.39	187.17	0.9503	24.89
	VAN	44.5	16.0	288	59.59	221.32	0.9398	25.20
	HorNet	45.7	16.3	287	55.79	202.73	0.9456	24.49
	MogaNet	46.8	16.5	255	49.48	184.11	0.9521	25.07

- Codebase: OpenSTL.

[1] OpenSTL: A Comprehensive Benchmark of Spatio-Temporal Predictive Learning. NeurIPS, 2023.

西湖大學
WESTLAKE UNIVERSITY

浙江大學
ZHEJIANG UNIVERSITY

ICLR
The 12th Internaltional Conference on Learning Representations

Thank you!

Paper: MogaNet

Source Code: MogaNet

Homepage

lisiyuan@westlake.edu.cn