

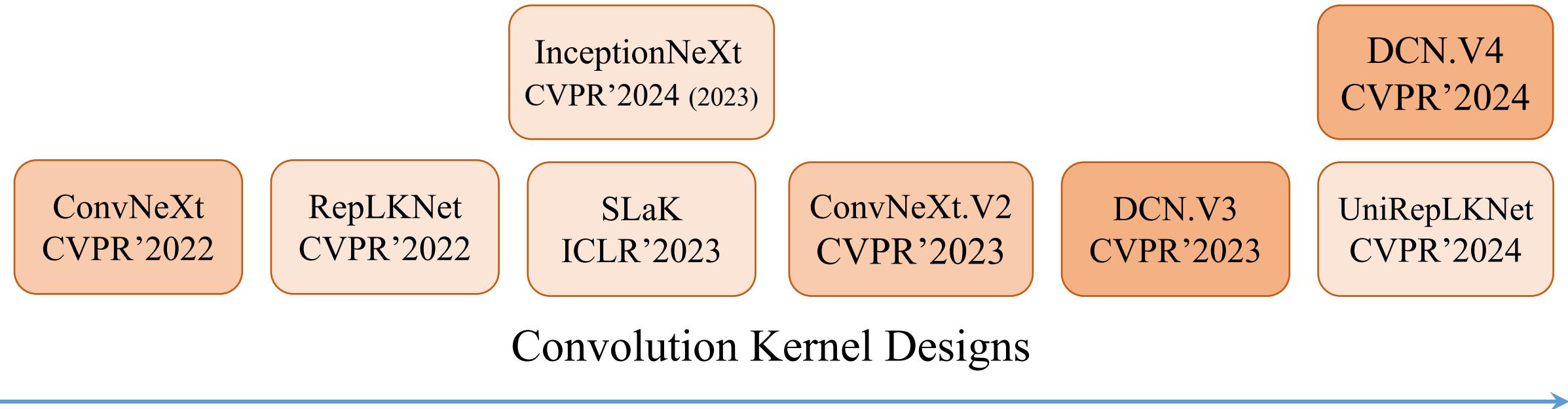


Modern Convolutional Neural Networks

Siyuan Li

Westlake University, Zhejiang University
March, 2024

Timeline of Modern CNNs



Content

1. Modern CNNs: Macro Design and Pre-training

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

2. Design of Convolution Kernels

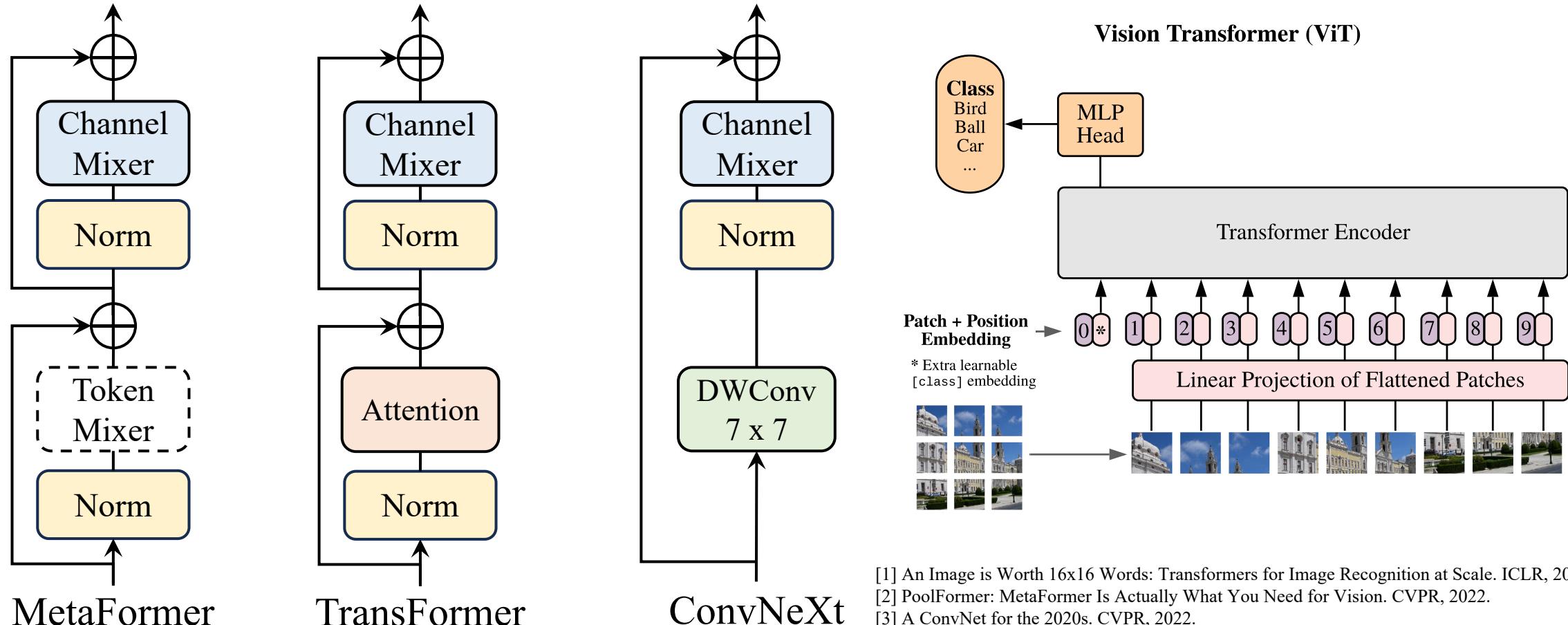
RepLKNet, SLaK, InceptionNext, DCN.V3/V4, UniRepLKNet

3. Combining Large Kernel with Gated Attention

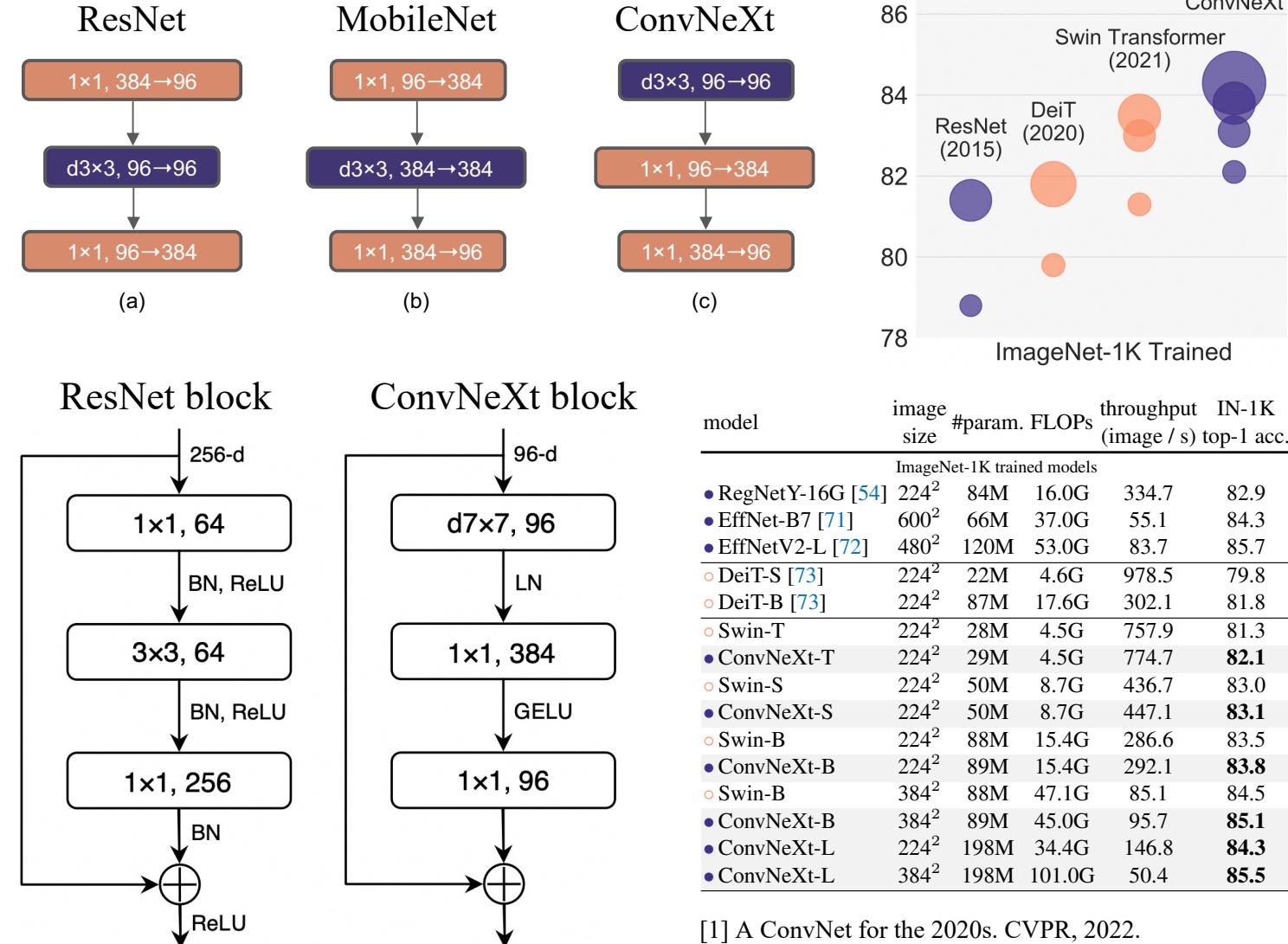
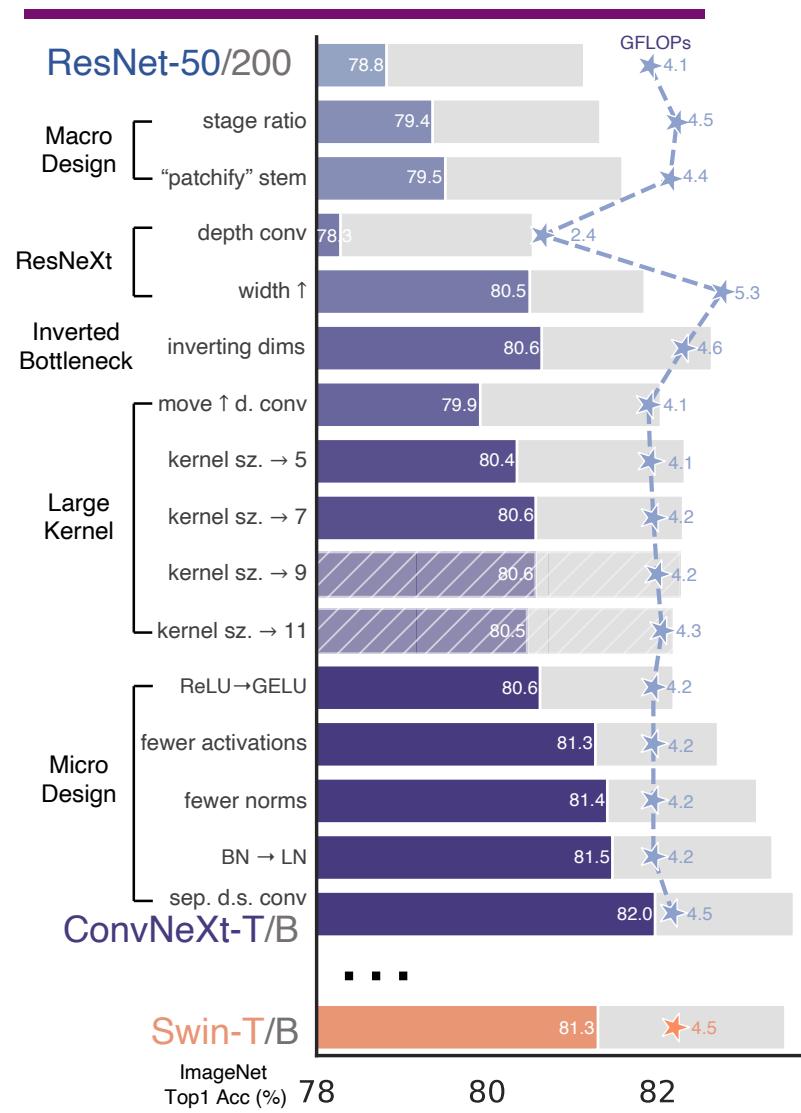
VAN, HorNet, FocalNet, MogaNet, Mamba, VMamba

Modern CNNs: Macro Design

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

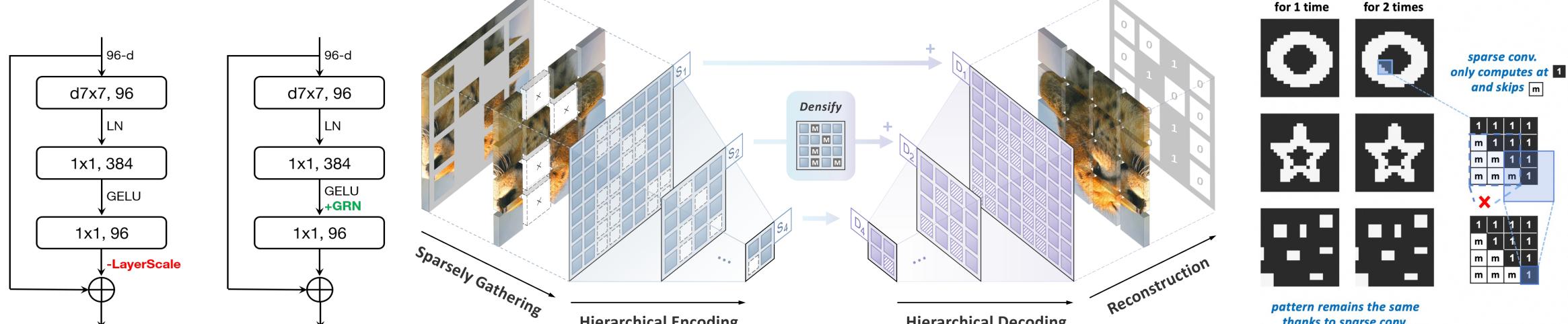


Modern CNNs: ConvNeXt



Modern CNNs: ConvNeXt.V2

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



ConvNeXt.V1 ConvNeXt.V2

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
```

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

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

[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.

| 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 | #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 |

Content

1. Modern CNNs: Macro Design and Pre-training

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

2. Design of Convolution Kernels

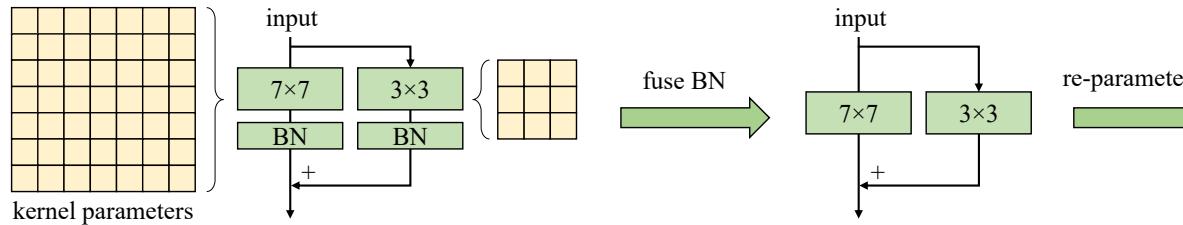
RepLKNet, SLaK, InceptionNext, DCN.V3/V4, UniRepLKNet

3. Combining Large Kernel with Gated Attention

VAN, HorNet, FocalNet, MogaNet, Mamba, VMamba

Large Kernels: RepLKNet

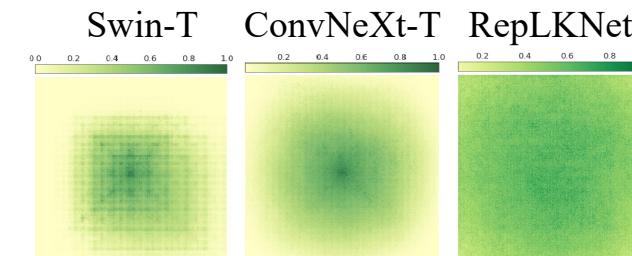
- Large-Kernel (LK) Convolutions are **efficient** and **competitive** as Self-attention.
- Training extremely large convolutions with **Structural Re-parameterization**.



| 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 |

| Kernel size | Architecture | ImageNet | | | ADE20K | | |
|-------------|----------------|----------|--------|-------|-------------|--------|-------|
| | | Top-1 | Params | FLOPs | 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 |

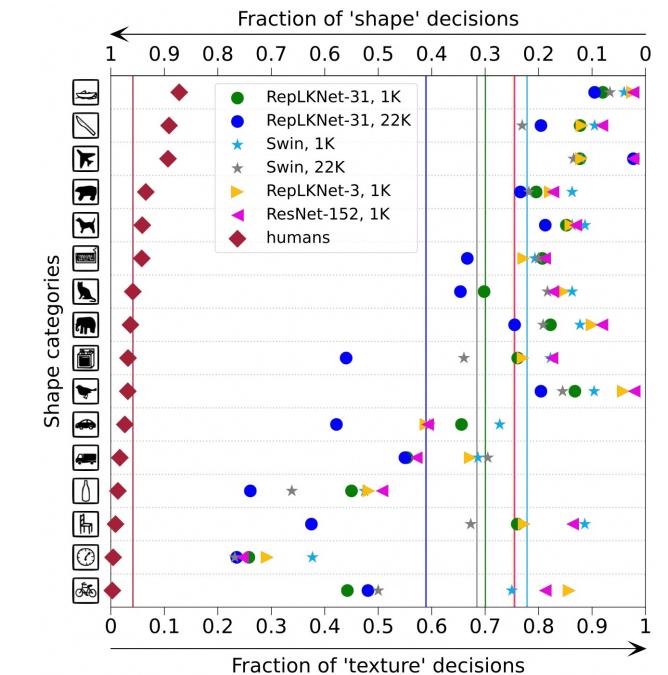
Extremely large kernels benefit both classification and downstream tasks and outperforms ViTs.



Effective receptive field

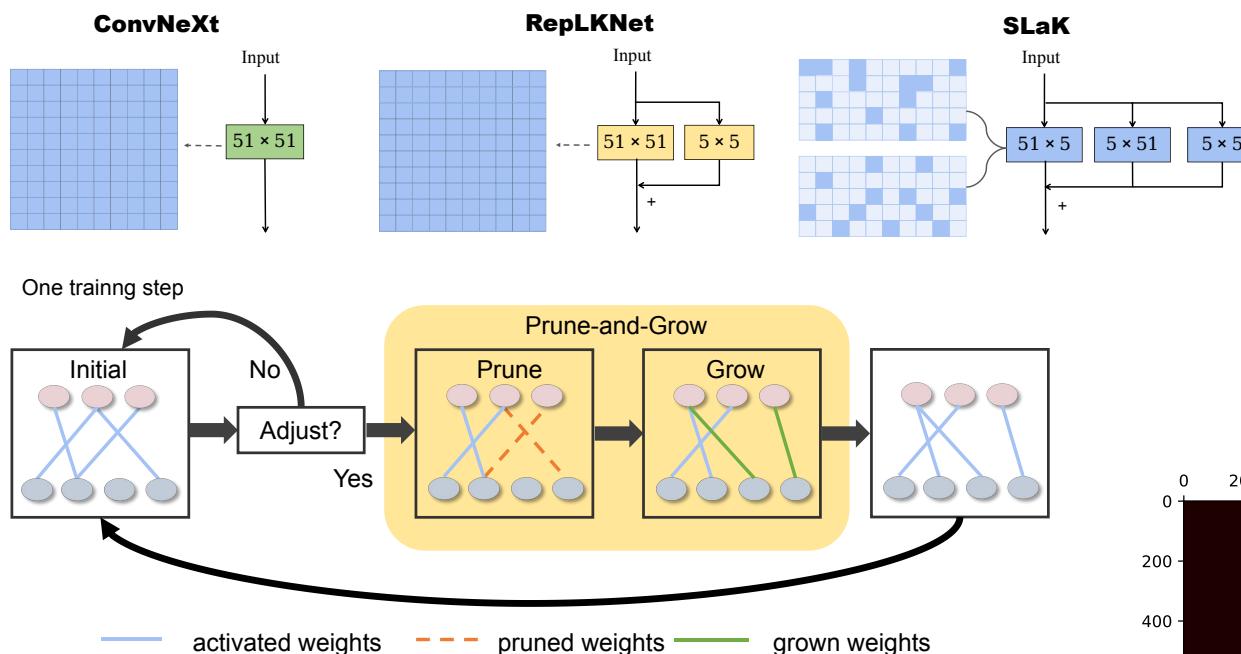
$$\text{DW7} \times 7 = \text{DW3} \times 3 (\text{BN}) + \text{DW7} \times 7 (\text{BN}) + \text{Short-cut.}$$

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



Large Kernels: SLaK

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



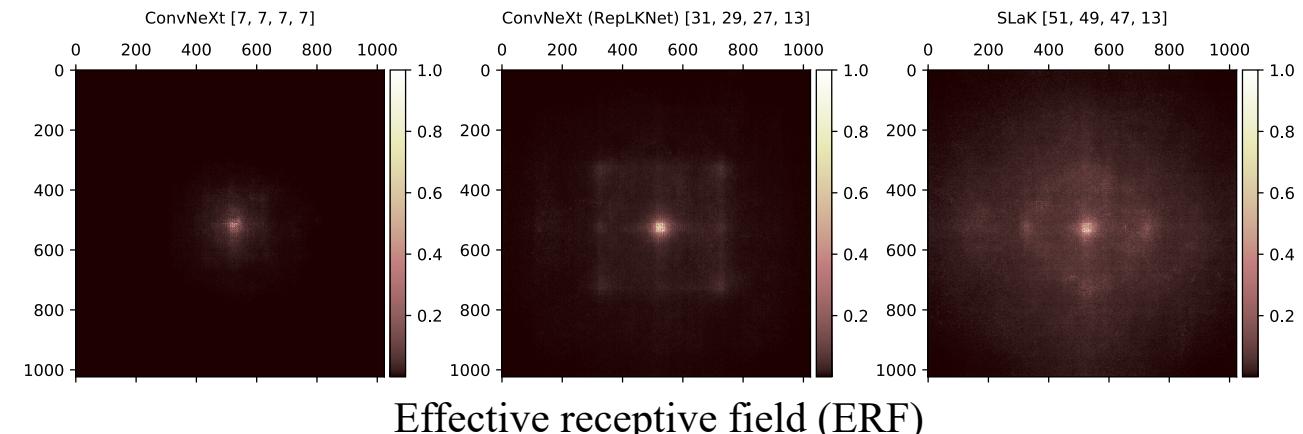
- (1) Initialization: Constructing Sparse Convolution based on SNIP^[2]
- (2) Dynamic sparsity: Pruning (the lowest magnitude) and growing

[1] More ConvNets in the 2020s: Scaling up Kernels Beyond 51×51 using Sparsity. ICLR, 2023.

[2] SNIP: Single-shot Network Pruning based on Connection Sensitivity. ICLR, 2019.

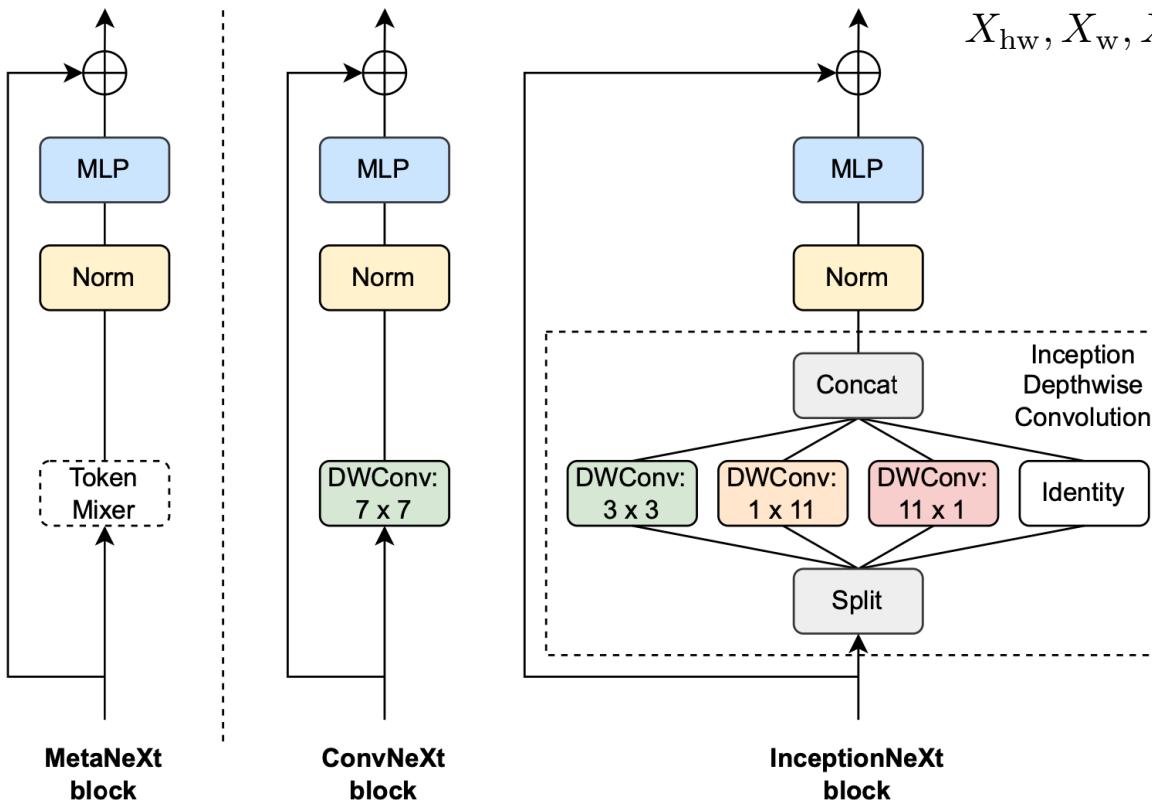
| Kernel Size | Top-1 Acc | #Params | FLOPs | Decomposed | | Sparse groups | | Sparse groups, expand more width | |
|-------------|-----------|---------|-------|------------|----------|---------------|----------|----------------------------------|----------|
| | | | | ConvNeXt-T | RepLKNet | ConvNeXt-T | RepLKNet | ConvNeXt-T | RepLKNet |
| 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 ^{mask} | | |
|---|-------------|---------------------------------|---------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| | | AP ₅₀ ^{box} | AP ₇₅ ^{box} | AP ₅₀ ^{mask} | AP ₇₅ ^{mask} | AP ₅₀ ^{mask} | AP ₇₅ ^{mask} |
| 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 |



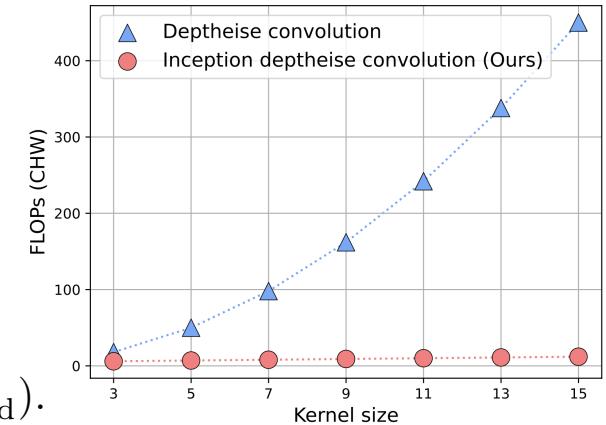
Large Kernels: InceptionNeXt

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



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

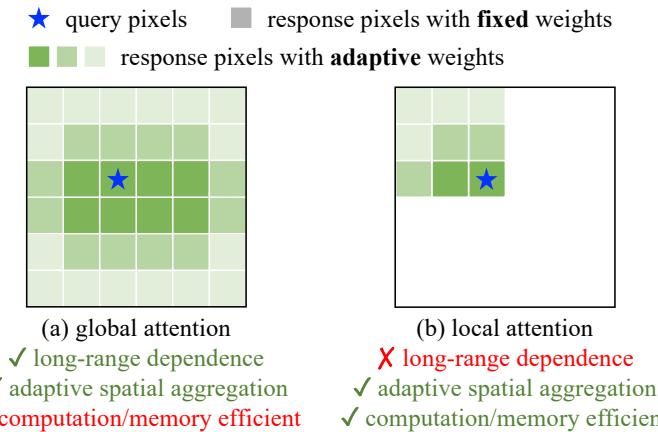
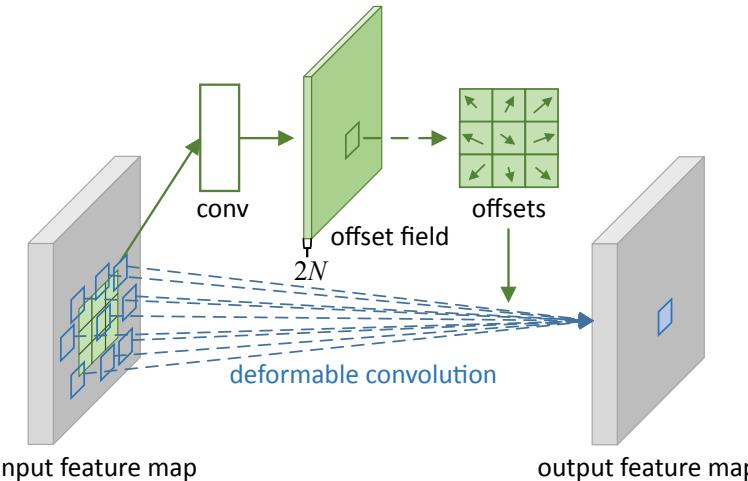
| | |
|---------------------------------------|--|
| Inception Depthwise Convolution | $X_{\text{hw}}, X_{\text{w}}, X_{\text{h}}, X_{\text{id}} = \text{Split}(X)$ $= X_{:, :, g}, X_{:, g: 2g}, X_{:, 2g: 3g}, X_{:, 3g:}$ $X'_{\text{hw}} = \text{DWConv}_{k_s \times k_s}^{g \rightarrow g} g(X_{\text{hw}}),$ $X'_{\text{w}} = \text{DWConv}_{1 \times k_b}^{g \rightarrow g} g(X_{\text{w}}),$ $X'_{\text{h}} = \text{DWConv}_{k_b \times 1}^{g \rightarrow g} g(X_{\text{h}}),$ $X'_{\text{id}} = X_{\text{id}}.$ $X' = \text{Concat}(X'_{\text{hw}}, X'_{\text{w}}, X'_{\text{h}}, X'_{\text{id}})$ |
|---------------------------------------|--|



| Model | Mixing Type | Image (size) | Params (M) | MACs (G) | Throughput (img/second) | | Top-1 (%) |
|------------------------|-------------|------------------|------------|----------|-------------------------|-------------|-------------|
| | | | | | Train | Inference | |
| 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) |

Kernel Designs: DCN.V3 (InternImage)

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



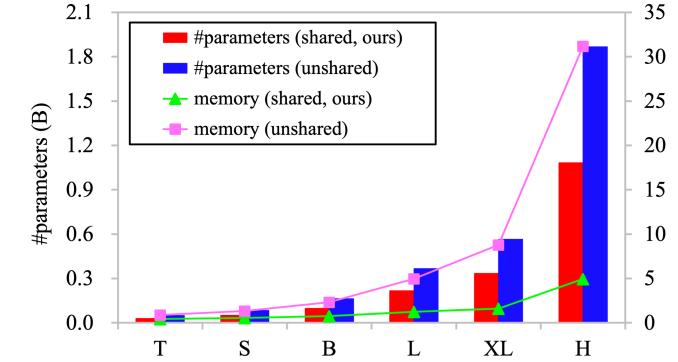
$$\text{DCN.V1: } \mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}}^K \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n)$$

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

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

Offsets $\Delta\mathbf{p}_n$, Regular grids \mathbf{p}_n , Modulation \mathbf{m}_k , weights \mathbf{w}

Scaling-up with efficient impl.

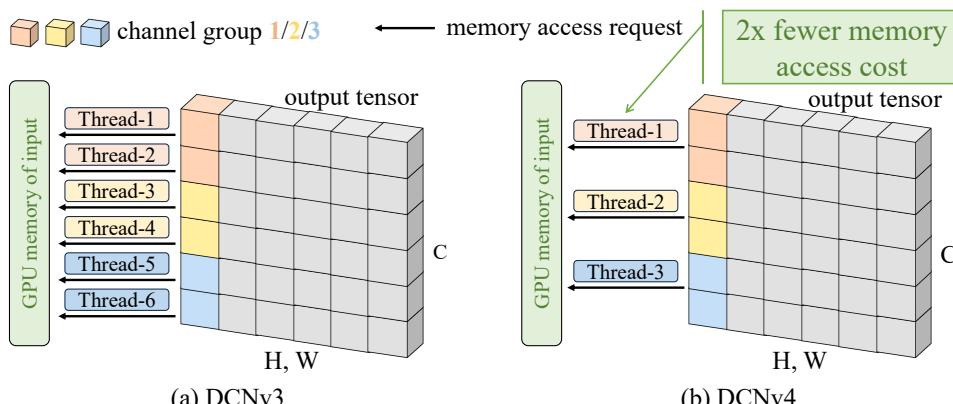
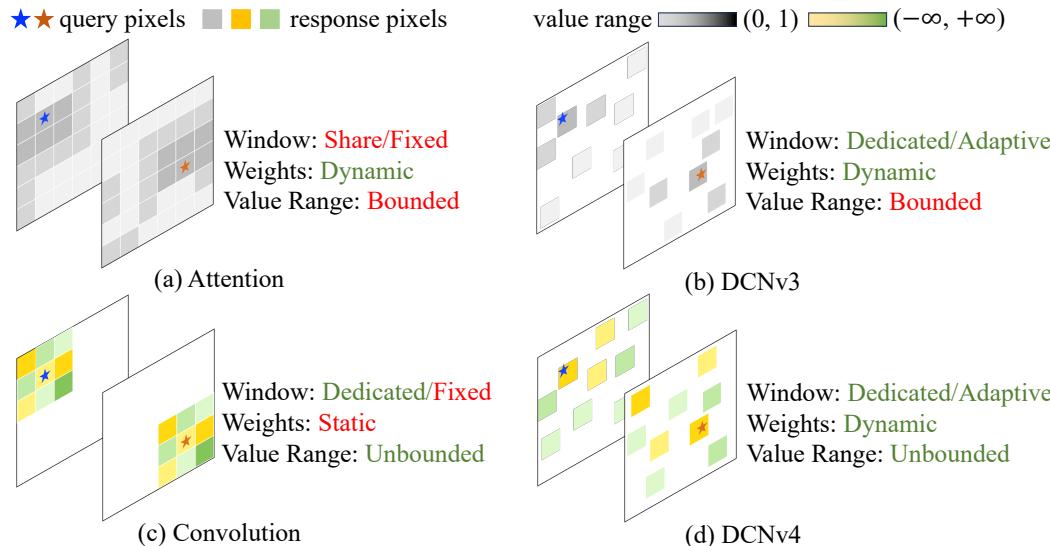


[1] Deformable Convolutional Networks. ICCV, 2017. [2] Deformable ConvNets v2: More Deformable, Better Results. CVPR, 2018.

[3] InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. CVPR, 2023.

Kernel Designs: DCN.V4 (FlashInternImage)

- 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 | 25.3 | 51.1 | 69.1 | 81.6 |
| | (-21.4) | (-28.2) | (-15.0) | (-5.7) | (-2.2) |

Using Softmax in DWConv 7×7 degrading 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 | 3 × +MS AP ^b |
| Swin-L | 253M | 20 / 26 | 51.8 | 44.9 |
| ConvNeXt-L | 255M | 26 / 40 | 53.5 | 46.4 |
| InternImage-L | 277M | 20 / 26 | 54.9 | 47.7 |
| ConvNeXt-XL | 407M | 21 / 32 | 53.6 | 46.5 |
| InternImage-XL | 387M | 16 / 23 | 55.3 | 48.1 |
| FlashInternImage-L | 277M | 26 / 39 | 55.6 | 48.2 |
| | | | 56.7 | 48.9 |

Content

1. Modern CNNs: Macro Design and Pre-training

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

2. Design of Convolution Kernels

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

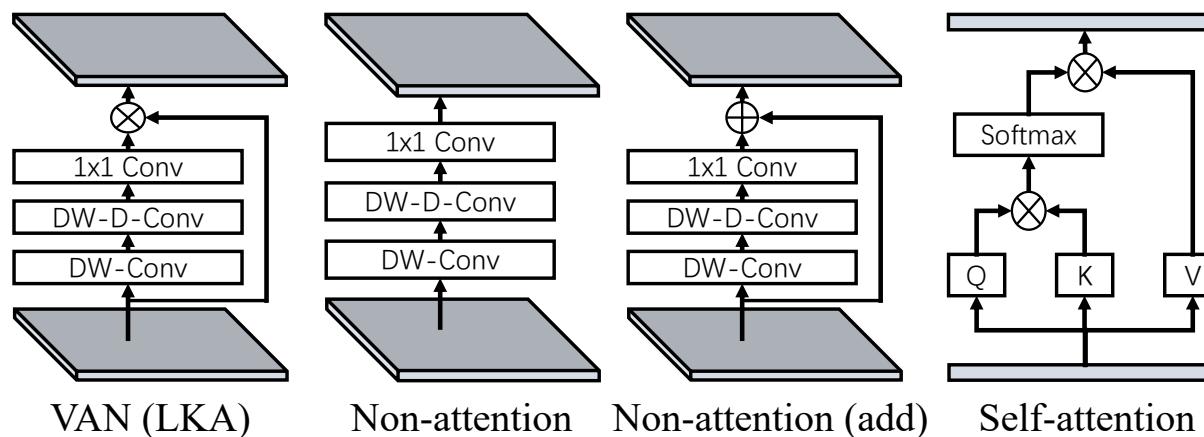
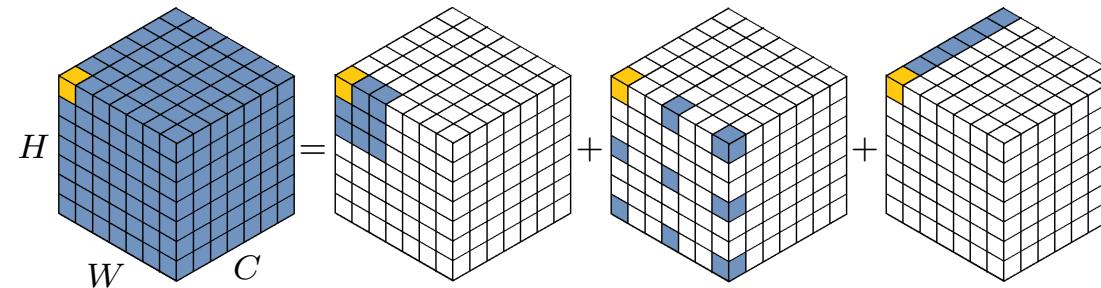
3. Combining Large Kernel with Gated Attention

VAN, HorNet, FocalNet, MogaNet, Mamba, VMamba

Gating & Large-kernel: VAN

- Decomposed large kernel + Gating.

$$\text{Conv}_{9 \times 9} = \text{DWConv}_{3 \times 3} + \text{DWConv}_{3 \times 3} + \text{PWConv}_{1 \times 1} \quad (\text{Dilation}=3)$$



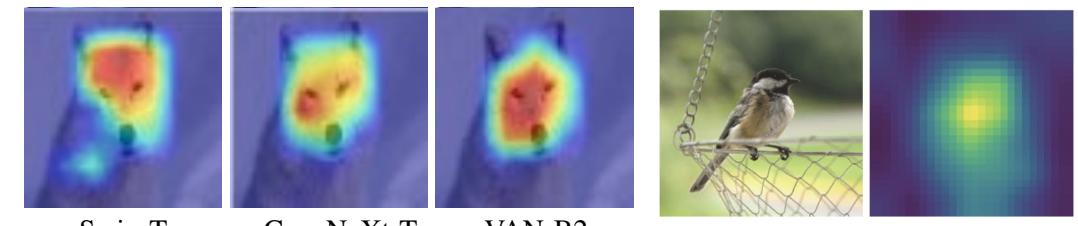
| Properties | Convolution | Self-Attention | LKA |
|--------------------------|------------------|--------------------|------------------|
| Local Receptive Field | ✓ | ✗ | ✓ |
| Long-range Dependence | ✗ | ✓ | ✓ |
| Spatial Adaptability | ✗ | ✓ | ✓ |
| Channel Adaptability | ✗ | ✗ | ✓ |
| Computational 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 (%)

$$\text{Conv}_{21 \times 21} = \text{DWConv}_{5 \times 5} + \text{DWConv}_{7 \times 7} + \text{PWConv}_{1 \times 1} \quad (\text{Dilation}=3)$$

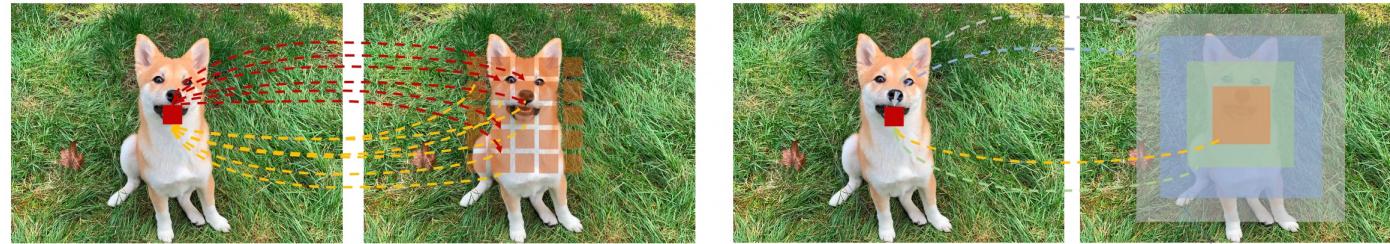


Grad-CAM visualization

Attention map visualization

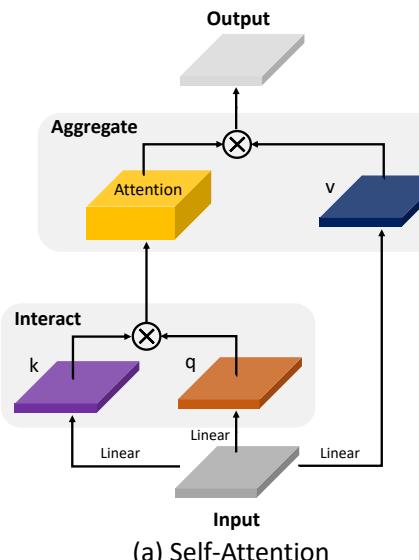
Gating & Hierarchical Kernel: FocalNet

- Hierarchical Contextualization + Gated Aggregation.



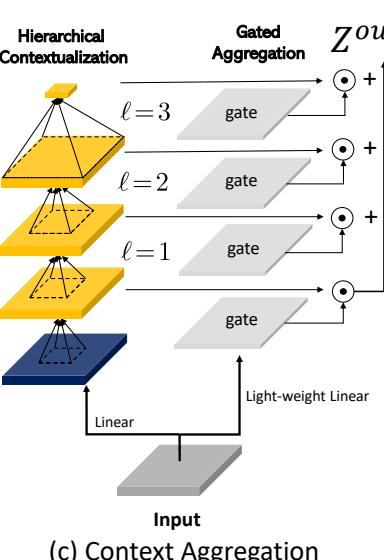
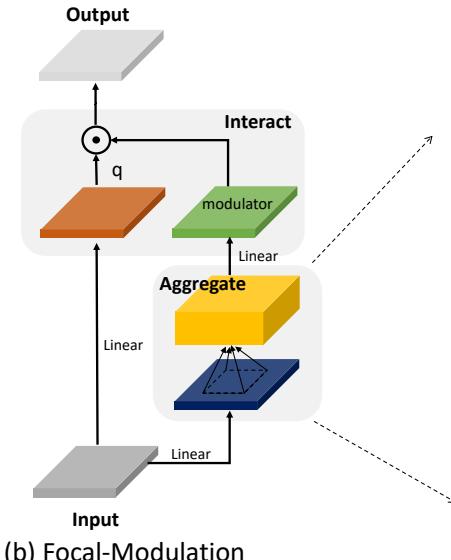
Self-Attention

- Query-Key Interaction
- Query-Value Aggregation



Focal Modulation

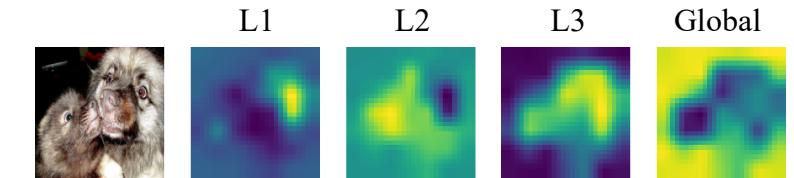
- Focal Aggregation
- Query-Modulator Interaction



```

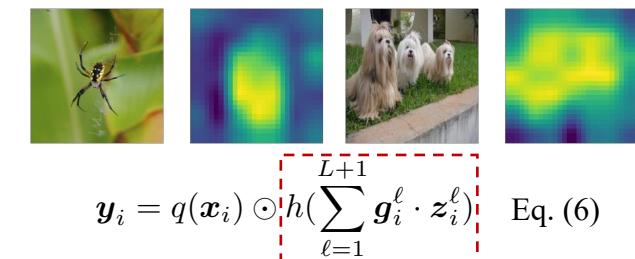
5 def forward(x, m=0):
6     x = pj_in(x).permute(0, 3, 1, 2)
7     q, z, gate = split(x, (C, C, L+1), 1)
8     for ℓ in range(L):
9         z = hc_layers[ℓ](z)           # Eq.(4), hierarchical contextualization
10        m = m + z * gate[:, ℓ:ℓ+1]  # Eq.(5), gated aggregation
11    m = m + GeLU(z.mean(dim=(2,3))) * gate[:, L:]
12    x = q * pj_cxt(m)            # Eq.(6), Focal Modulation
13    return pj_out(x.permute(0, 2, 3, 1))

```



$$\mathbf{Z}^\ell = f_a^\ell(\mathbf{Z}^{\ell-1}) \triangleq \text{GeLU}(\text{DWConv}(\mathbf{Z}^{\ell-1})) \in \mathbb{R}^{H \times W \times C} \quad \text{Eq. (4)}$$

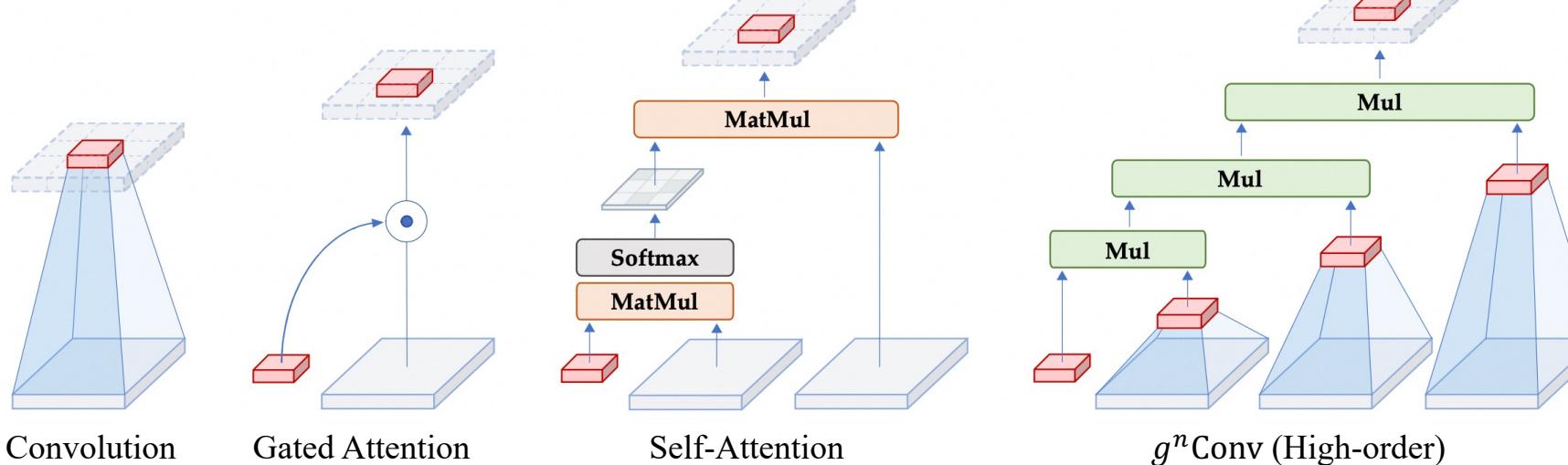
$$\mathbf{Z}^{out} = \sum_{\ell=1}^{L+1} \mathbf{G}^\ell \odot \mathbf{Z}^\ell \in \mathbb{R}^{H \times W \times C} \quad \text{Eq. (5)}$$



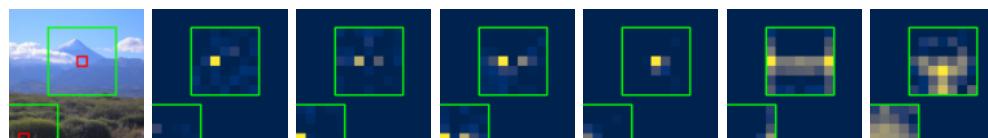
$$\mathbf{y}_i = q(\mathbf{x}_i) \odot h\left(\sum_{\ell=1}^{L+1} \mathbf{g}_i^\ell \cdot \mathbf{z}_i^\ell\right) \quad \text{Eq. (6)}$$

Gating & Hierarchical Kernel: HorNet

- High-order Interactions: Recursive DWConv + Gating.

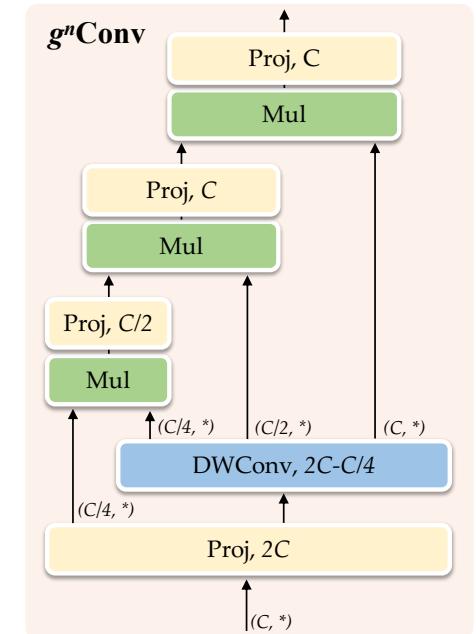


$$x_{g^n\text{Conv}}^{(i,c)} = p_n^{(i,c)} = \sum_{j \in \Omega_i} \sum_{c'=1}^C w_{n-1,i \rightarrow j}^c \mathbf{g}_{n-1}^{(i,c)} w_{\phi_{\text{in}}}^{(c',c)} x^{(j,c')} \triangleq \sum_{j \in \Omega_i} \sum_{c'=1}^C h_{ij}^c w_{\phi_{\text{in}}}^{(c',c)} x^{(j,c')} \quad \text{Eq. (3.8)}$$



Adaptive weights generated by $g^n\text{Conv}$, i.e., $\frac{1}{C} \sum_{c=1}^C h_{ij}^c$ in Eq. (3.8)

[1] HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. NeurIPS, 2022.



```

def forward(self, x):
    x = self.proj_in(x)
    y, x = torch.split(x, (self.dims[0], sum(self.dims)), dim=1)
    x = self.dwconv(x)
    x_list = torch.split(x, self.dims, dim=1)
    x = y * x_list[0]
    for i in range(self.order - 1):
        x = self.projs[i](x) * x_list[i+1]
    return self.proj_out(x)

self.projs = nn.ModuleList([
    nn.Conv2d(self.dims[i], self.dims[i+1], 1)
    for i in range(order-1)])
self.proj_out = nn.Conv2d(dim, dim, 1)
  
```

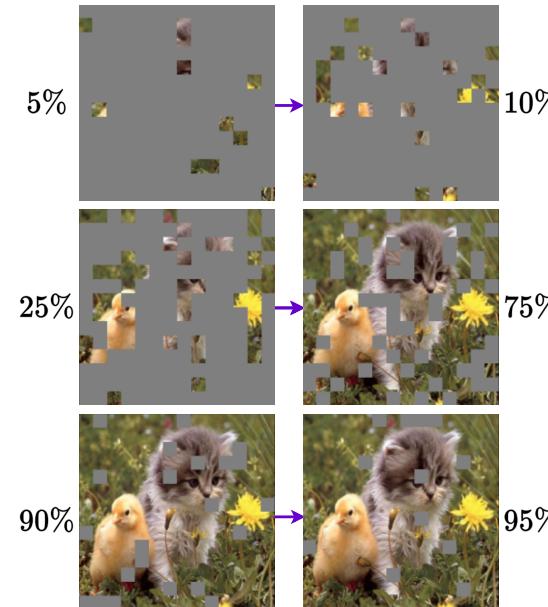
Multi-order Interaction: MogaNet

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

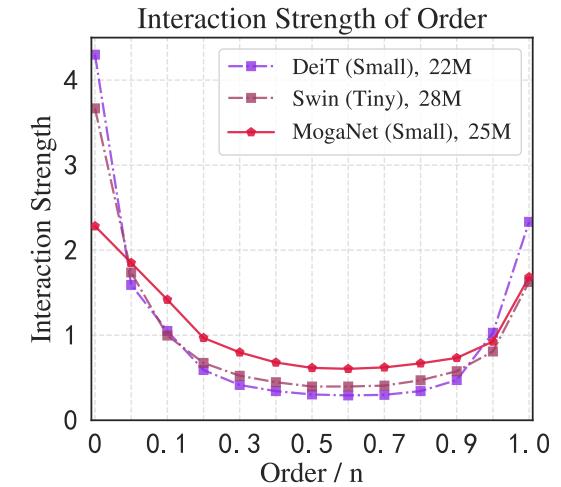
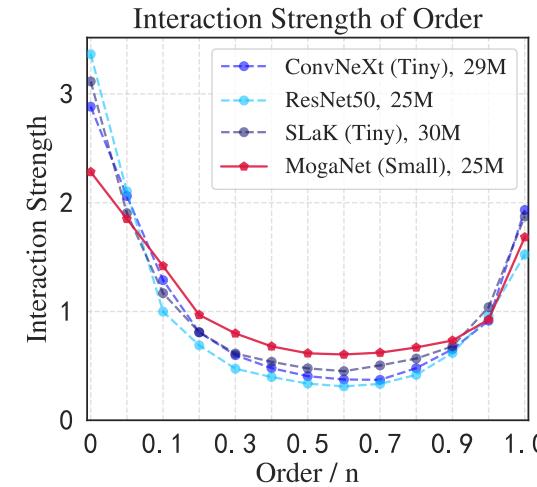
Multi-order Interactions $I^{(m)}(i, j) = \mathbb{E}_{S \subseteq N \setminus \{i, j\}, |S|=m} [\Delta f(i, j, S)]$
 $N = \{1, \dots, n\}$ $0 \leq m \geq n - 2$
 $\Delta f(i, j, S) = f(S \cup \{i, j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S)$

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

-  **Much** new information
-  **Little** new infomation
-  **Little** new information
-  **Much** new infomation
-  **Much** new information
-  **Little** new infomation

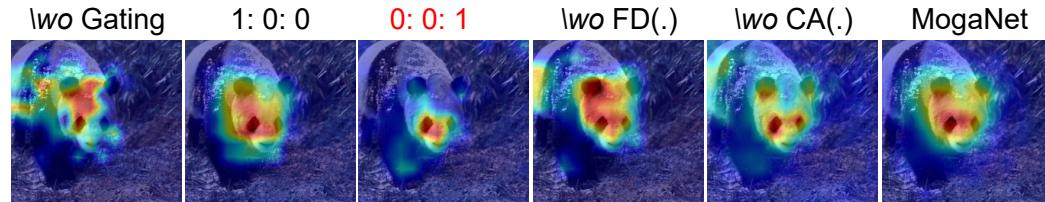
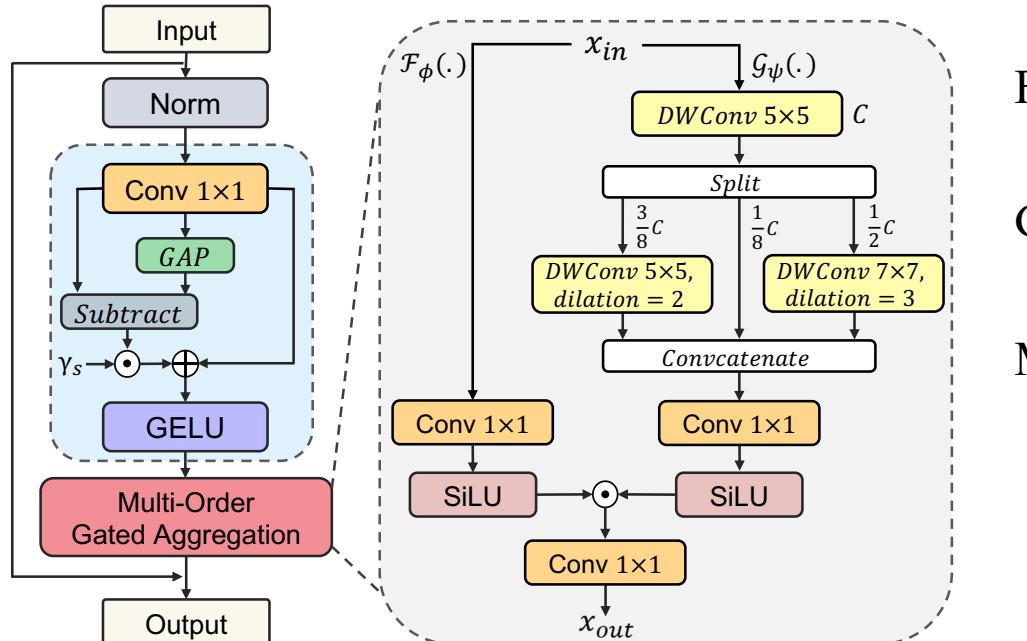


Both ViTs and modern CNN architectures fail to explore middle-order interactions, which are informative to humans.



Multi-order Interaction: MogaNet

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



Feature decomposition:

$$Y = \text{Conv}_{1 \times 1}(X),$$

$$Z = \text{GELU}\left(Y + \gamma_s \odot (Y - \text{GAP}(Y))\right)$$

Gated aggregation branch: $Z = \underbrace{\text{SiLU}(\text{Conv}_{1 \times 1}(X))}_{\mathcal{F}_\phi} \odot \underbrace{\text{SiLU}(\text{Conv}_{1 \times 1}(Y_C))}_{\mathcal{G}_\psi}$

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

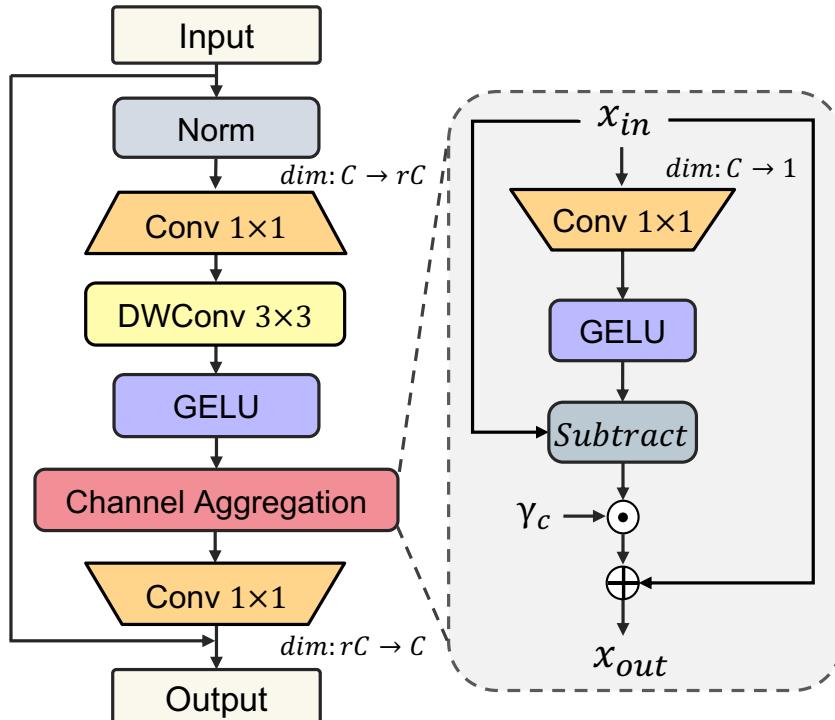
$$C_l + C_m + C_h = C, Y_C = \text{Concat}(Y_{l,1:C_l}, Y_m, Y_h)$$

| Modules | Top-1 Acc (%) | Params. (M) | FLOPs (G) | Top-1 Acc (%) | | Context branch | | |
|---|---------------|-------------|-----------|---------------|------|----------------|------|------|
| | | | | None | GELU | SiLU | None | GELU |
| Baseline (+Gating branch) | 77.2 | 5.09 | 1.070 | 76.3 | 76.7 | 76.7 | 76.3 | 76.7 |
| DW _{7×7} | 77.4 | 5.14 | 1.094 | 76.8 | 77.0 | 76.9 | 76.8 | 77.0 |
| DW _{5×5,d=1} + DW _{7×7,d=3} | 77.5 | 5.15 | 1.112 | 76.7 | 76.8 | 77.0 | 76.7 | 77.0 |
| DW _{5×5,d=1} + DW _{5×5,d=2} + DW _{7×7,d=3} | 77.5 | 5.17 | 1.185 | 76.9 | 77.1 | 77.2 | 76.9 | 77.1 |
| +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 | | | | | |

Ablation of SA module with MogaNet-T on ImageNet

Multi-order Interaction: MogaNet

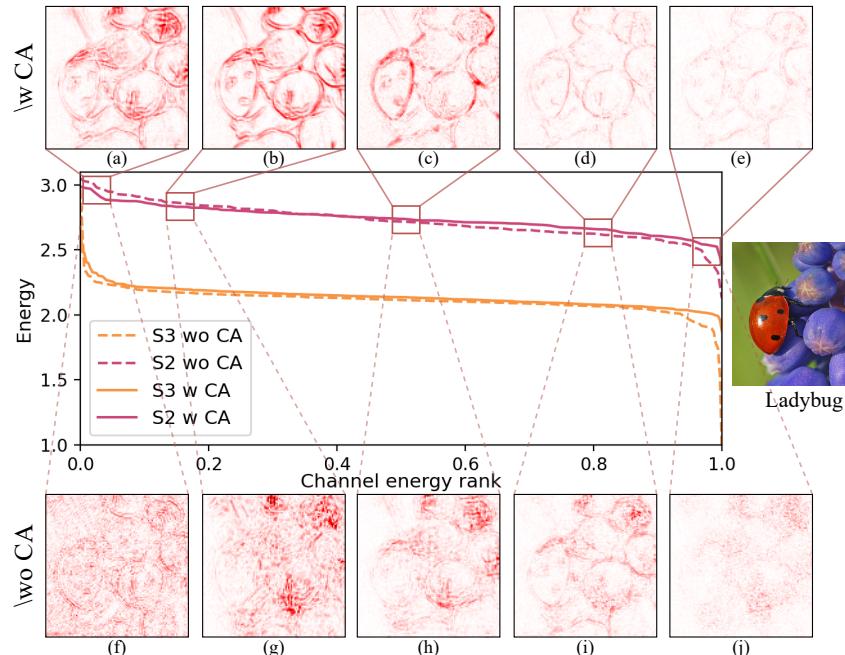
- Channel Aggregation (CA): Multi-order Channel Reallocation.



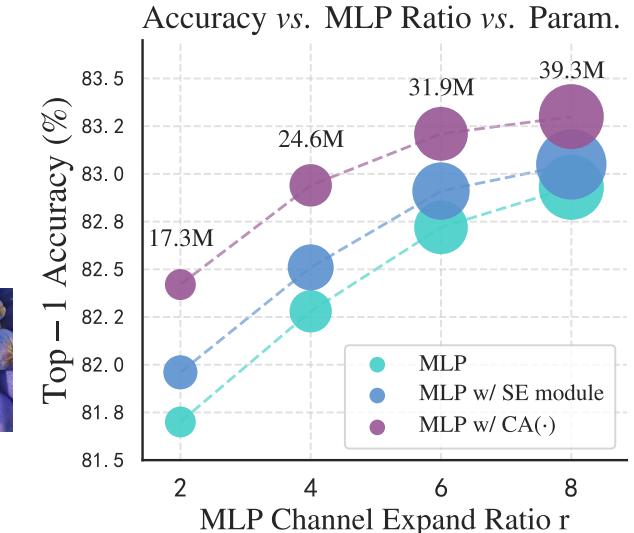
$$Y = \text{GELU}\left(\text{DW}_{3 \times 3}\left(\text{Conv}_{1 \times 1}(\text{Norm}(X))\right)\right),$$

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

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



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



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

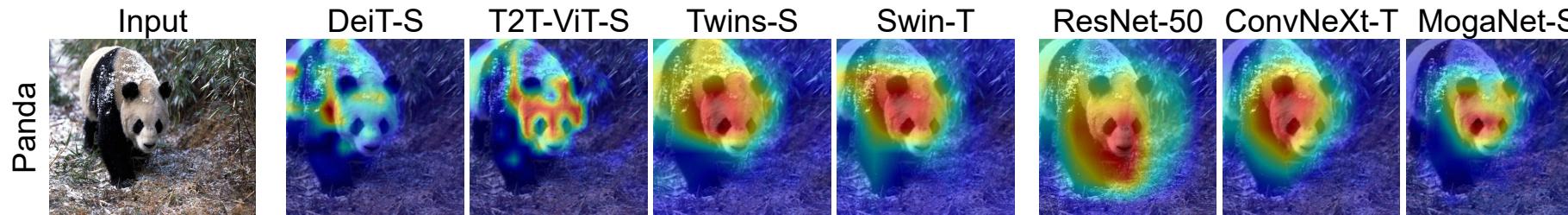
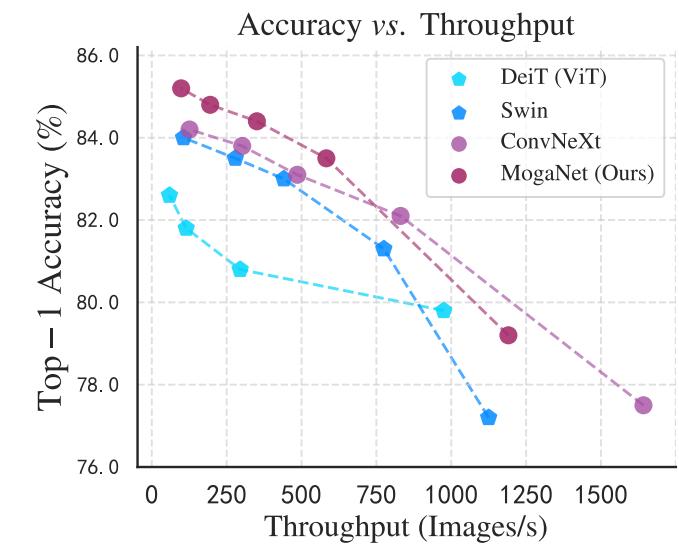
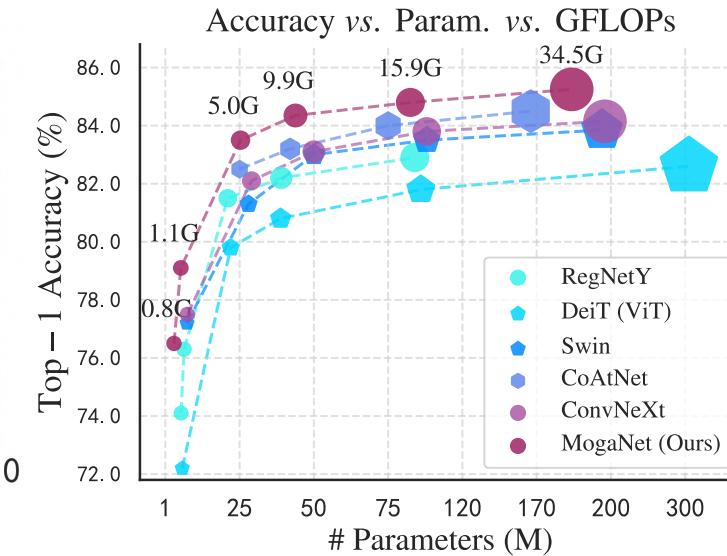
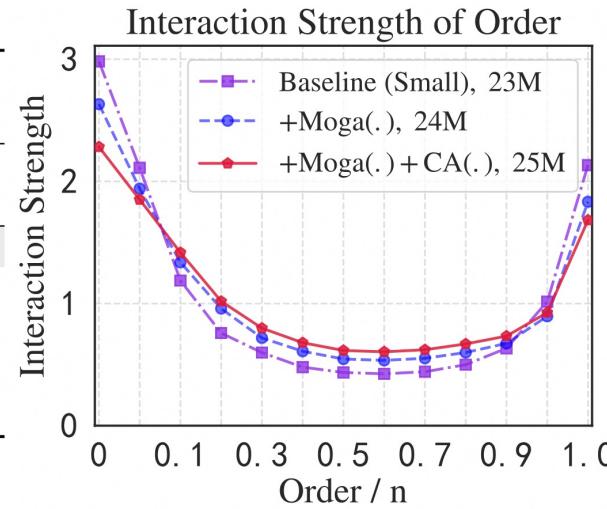
Ablation of MogaNet-S on ImageNet

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

Multi-order Interaction: MogaNet

- Great scalability and efficiency of parameters.
- Relieving representation bottleneck.

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



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 | X | X | 76.5 |
| MogaNet-XT | 224 ² | 2×10^{-3} | 20 | X | ✓ | X | 77.1 |
| MogaNet-XT | 256 ² | 1×10^{-3} | 5 | 7/0.5 | X | X | 77.2 |
| MogaNet-XT | 256 ² | 2×10^{-3} | 20 | X | ✓ | X | 77.6 |
| MogaNet-T | 224 ² | 1×10^{-3} | 5 | 7/0.5 | X | X | 79.0 |
| MogaNet-T | 224 ² | 2×10^{-3} | 20 | X | ✓ | X | 79.4 |
| MogaNet-T | 256 ² | 1×10^{-3} | 5 | 7/0.5 | X | X | 79.6 |
| MogaNet-T | 256 ² | 2×10^{-3} | 20 | X | ✓ | X | 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 _{7×7} | 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 _{7×7} | 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 |

Large size (80-200M)

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

Training and inference at the resolution of 224² or 256².

MogaNet: COCO Object Det. and Seg.

RetinaNet (1×)

| Architecture | Type | #P. | FLOPs | RetinaNet 1× | | | | | | | |
|-------------------|------|-----|-------|--------------|-------------|-------------|------------------|------------------|-----------------|-----------------|-----------------|
| | | | | (M) | (G) | 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. | FLOPs | Mask R-CNN 1× | | | | | | | |
|-------------------|------|-----|-------|---------------|-------------|-----------------|-------------------------------|-------------------------------|-----------------|-------------------------------|-------------------------------|
| | | | | (M) | (G) | AP ^b | AP ₅₀ ^b | AP ₇₅ ^b | AP ^m | AP ₅₀ ^m | AP ₇₅ ^m |
| 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. | FLOPs | Cascade Mask R-CNN +MS 3× | | | | | | | |
|-------------------------------|------|-----|-------|---------------------------|-------------|------------------|-------------------------------|-------------------------------|-----------------|-------------------------------|-------------------------------|
| | | | | (M) | (G) | AP ^{bb} | AP ₅₀ ^b | AP ₇₅ ^b | AP ^m | AP ₅₀ ^m | AP ₇₅ ^m |
| 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.

MogaNet: ADE20K Semantic Segmentation

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 resolutions.
- UperNet (160K) with 512×2048 or 640×2560 inference resolutions using IN-1K or IN-21K models.

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 |
| Twins-P-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 _{7×7} | 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 Pose Estimation

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 |

| 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 |

| Architecture | Type | Hand | | | Face | | | ↓ |
|------------------|------|------------|--------------|-------------------|------------|--------------|-------------|---|
| | | #P. (M) | FLOPs (G) | PA-MPJPE (mm)↓ | #P. (M) | FLOPs (G) | 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 | |

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

3D Human Pose with Expose

- 3D Face: FFHQ (256²)
- 3D Hand: FreiHand (224²)

MogaNet: Video Prediction

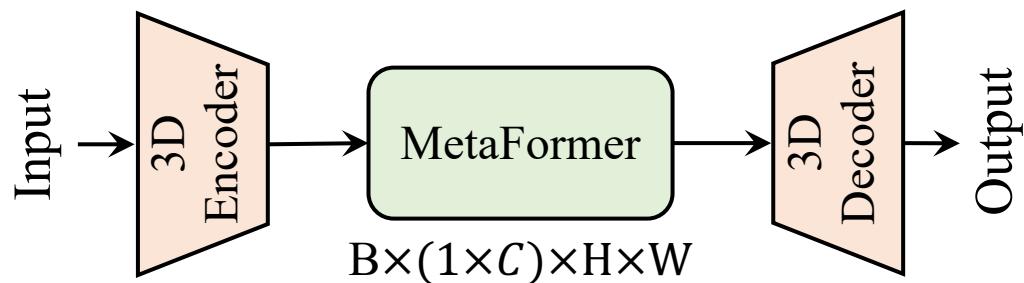
Moving MNIST ($10 \times 1 \times 64 \times 64$)

| Architecture | #P. (M) | FLOPs (G) | FPS (s) | 200 epochs | | | 2000 epochs | | |
|----------------|-------------|--------------|------------|--------------|--------------|---------------|--------------|--------------|---------------|
| | | | | MSE↓ | MAE↓ | SSIM↑ | MSE↓ | MAE↓ | 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 |

MMNIST-CIFAR ($10 \times 3 \times 64 \times 64$)

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

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



State-Space Models

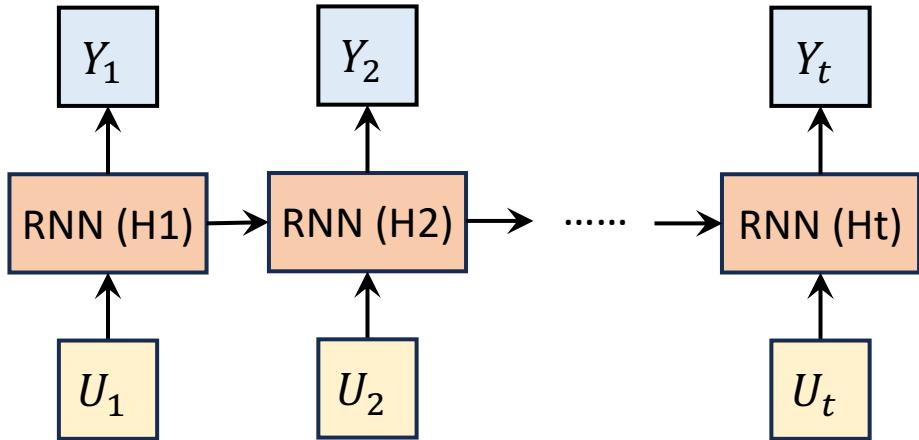
- State-Space Model (SSM): “Parallel RNN”
- SSM vs. Convolution: “Long Convolution”

SSM: $\hat{h}(t) = \mathbf{A}h(t) + \mathbf{B}u(t), \quad y(t) = \mathbf{C}h(t) + \mathbf{D}u(t).$

$$RNN: \begin{aligned} h_t &= \sigma(W_1 U_t + W_2 h_{t-1}), \\ o_t &= \sigma(W_3 h_t). \end{aligned}$$

$$y_k = \overline{\mathbf{C}\mathbf{A}}^k \overline{\mathbf{B}} u_0 + \overline{\mathbf{C}\mathbf{A}}^{k-1} \overline{\mathbf{B}} u_1 + \cdots + \overline{\mathbf{C}\mathbf{A}\mathbf{B}} u_{k-1} + \overline{\mathbf{C}\mathbf{B}} u_k$$

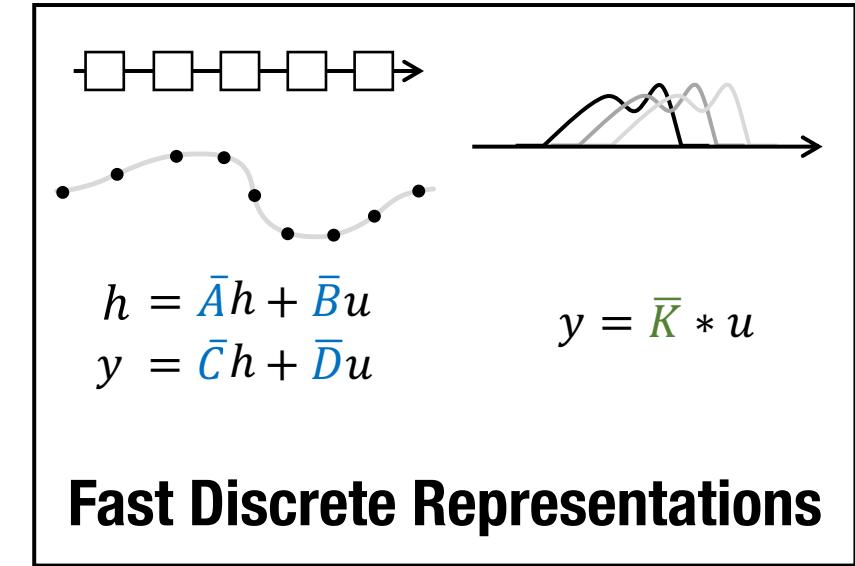
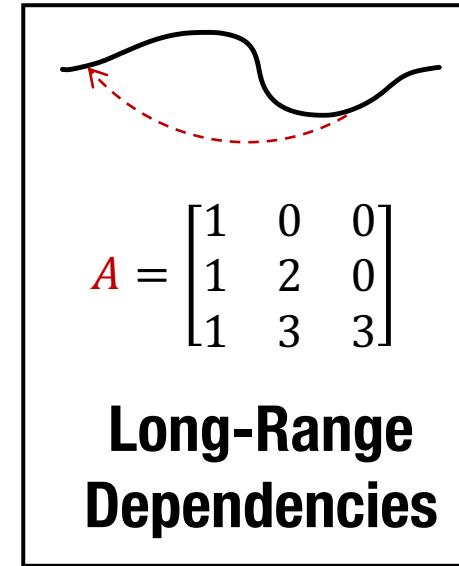
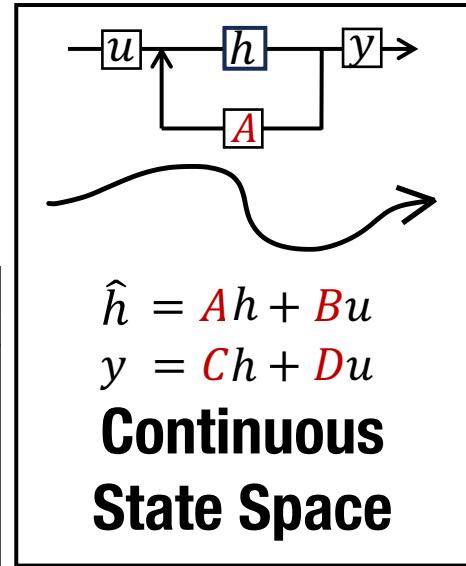
$$y = \overline{\mathbf{K}} * u.$$



HiPPO Matrix

$$\mathbf{A}_{nk} = \begin{cases} (-1)^{n-k}(2k+1) & n > k \\ k+1 & n = k \\ 0 & n < k \end{cases}$$

$$\mathbf{A} = \left[\begin{array}{ccccccc} 1 & 2 & & & & & \\ -1 & 3 & 3 & & & & \\ 1 & -3 & -5 & 4 & & & \\ -1 & 3 & -5 & -7 & 5 & & \\ 1 & -3 & 5 & -7 & -9 & 6 & \\ -1 & 3 & -5 & 7 & 9 & -11 & 7 \\ 1 & -3 & 5 & -7 & 9 & -11 & 7 \\ -1 & 3 & -5 & 7 & -9 & 11 & -13 & 8 \\ \vdots & & & & & & & \ddots \end{array} \right]$$



State-Space Models: Mamba

Structured state space sequence models (S4)

$$h'(t) = Ah(t) + Bx(t) \quad (1a)$$

$$y(t) = Ch(t) \quad (1b)$$

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \quad (2a)$$

$$y_t = Ch_t \quad (2b)$$

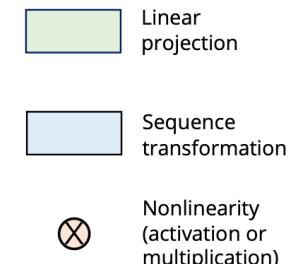
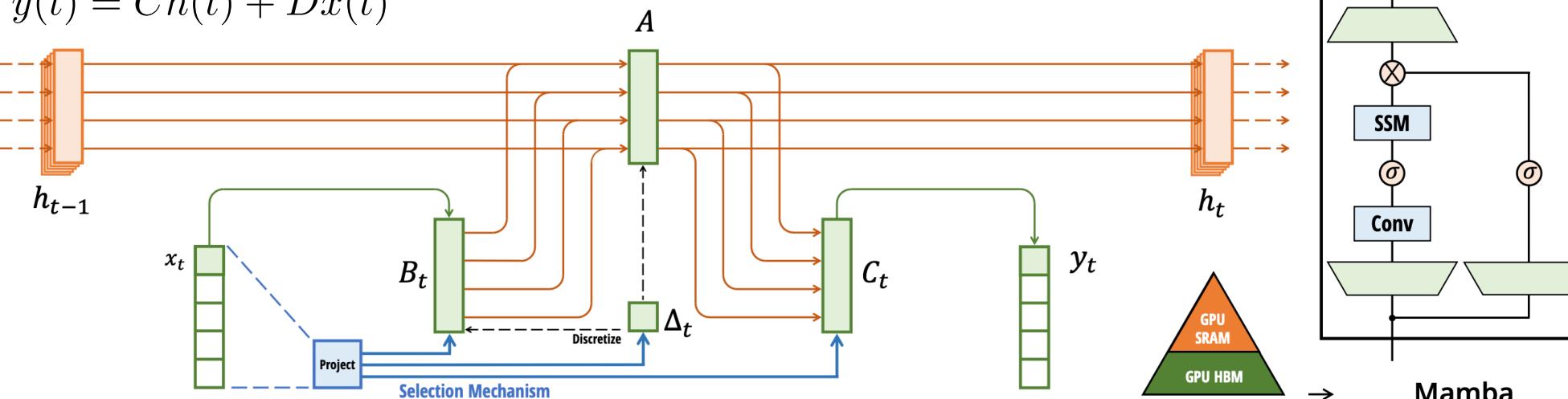
$$x(t) \in \mathbb{R}^L \rightarrow y(t) \in \mathbb{R}^L, A \in \mathbb{C}^{N \times N}, B, C \in \mathbb{C}^N, D \in \mathbb{C}^1$$

$$h'(t) = Ah(t) + Bx(t)$$

$$y(t) = Ch(t) + Dx(t)$$

$$\bar{\mathbf{K}} = (C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^{k-1}\bar{B}, \dots) \quad (3a)$$

$$y = x * \bar{\mathbf{K}} \quad (3b)$$



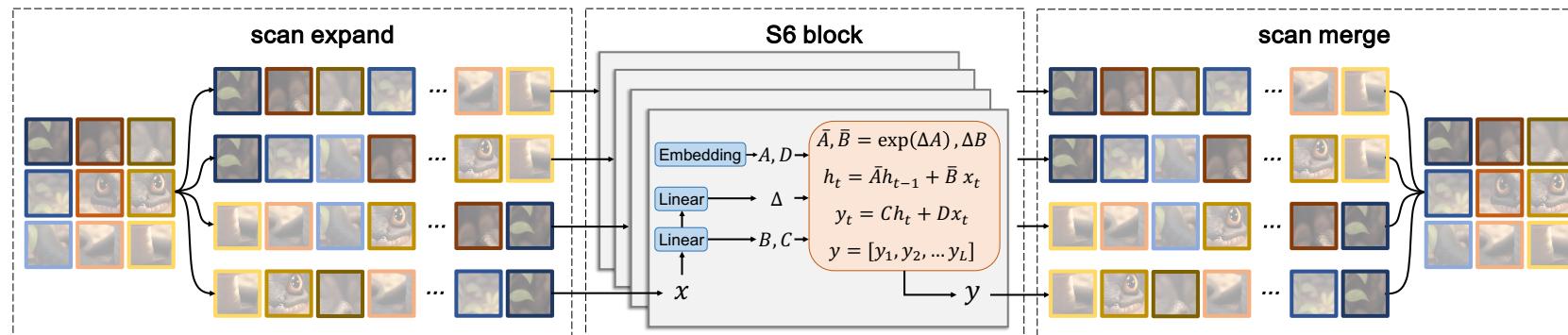
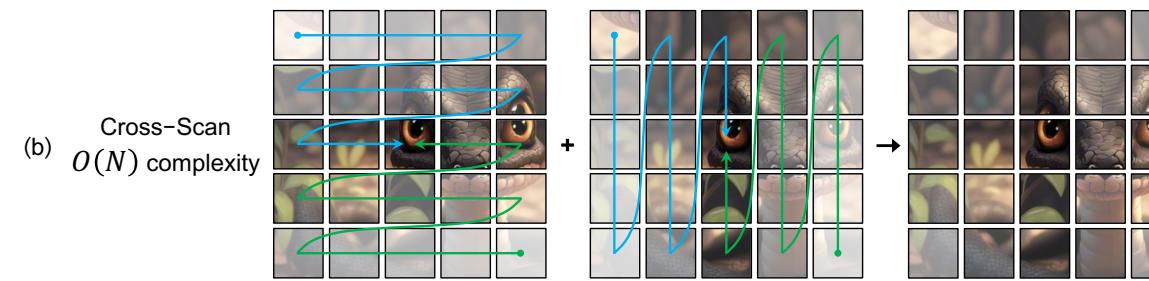
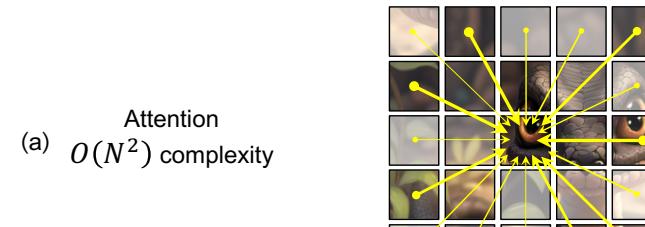
$$g_t = \sigma(\text{Linear}(x_t))$$

$$h_t = (1 - g_t)h_{t-1} + g_t x_t$$

| Model | Params | Accuracy (%) at Sequence Length | | | | | |
|----------|--------|---------------------------------|----------|----------|----------|----------|--------------|
| | | 2^{10} | 2^{12} | 2^{14} | 2^{16} | 2^{18} | 2^{20} |
| HyenaDNA | 1.4M | 28.04 | 28.43 | 41.17 | 42.22 | 31.10 | 54.87 |
| Mamba | 1.4M | 31.47 | 27.50 | 27.66 | 40.72 | 42.41 | 71.67 |
| Mamba | 7M | 30.00 | 29.01 | 31.48 | 43.73 | 56.60 | 81.31 |

Great Apes DNA Classification

State-Space Models: VMamba



ADE20K Segmentation

| method | crop size | mIoU (SS) | mIoU (MS) | #param. | FLOPs |
|--------------|-----------|-----------|-----------|---------|-------|
| ResNet-50 | 512^2 | 42.1 | 42.8 | 67M | 953G |
| DeiT-S + MLN | 512^2 | 43.8 | 45.1 | 58M | 1217G |
| Swin-T | 512^2 | 44.4 | 45.8 | 60M | 945G |
| ConvNeXt-T | 512^2 | 46.0 | 46.7 | 60M | 939G |
| VMamba-T | 512^2 | 47.3 | 48.3 | 55M | 939G |

ImageNet-1K Classification

| method | image size | #param. | FLOPs | ImageNet top-1 acc. |
|------------------|------------|---------|--------|---------------------|
| RegNetY-4G [36] | 224^2 | 21M | 4.0G | 80.0 |
| RegNetY-8G [36] | 224^2 | 39M | 8.0G | 81.7 |
| RegNetY-16G [36] | 224^2 | 84M | 16.0G | 82.9 |
| EffNet-B3 [42] | 300^2 | 12M | 1.8G | 81.6 |
| EffNet-B4 [42] | 380^2 | 19M | 4.2G | 82.9 |
| EffNet-B5 [42] | 456^2 | 30M | 9.9G | 83.6 |
| EffNet-B6 [42] | 528^2 | 43M | 19.0G | 84.0 |
| ViT-B/16 [10] | 384^2 | 86M | 55.4G | 77.9 |
| ViT-L/16 [10] | 384^2 | 307M | 190.7G | 76.5 |
| DeiT-S [45] | 224^2 | 22M | 4.6G | 79.8 |
| DeiT-B [45] | 224^2 | 86M | 17.5G | 81.8 |
| DeiT-B [45] | 384^2 | 86M | 55.4G | 83.1 |
| Swin-T [28] | 224^2 | 29M | 4.5G | 81.3 |
| Swin-S [28] | 224^2 | 50M | 8.7G | 83.0 |
| Swin-B [28] | 224^2 | 88M | 15.4G | 83.5 |
| S4ND-ViT-B [35] | 224^2 | 89M | - | 80.4 |
| VMamba-T | 224^2 | 22M | 4.5G | 82.2 |
| VMamba-S | 224^2 | 44M | 9.1G | 83.5 |

Thank you!



Paper: MogaNet



Code: MogaNet



Homepage



lisiyuan@westlake.edu.cn