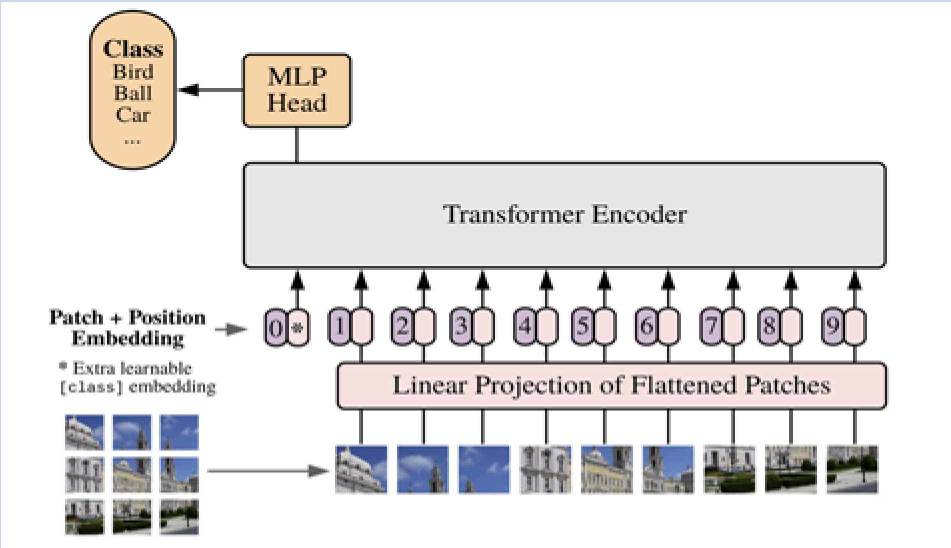
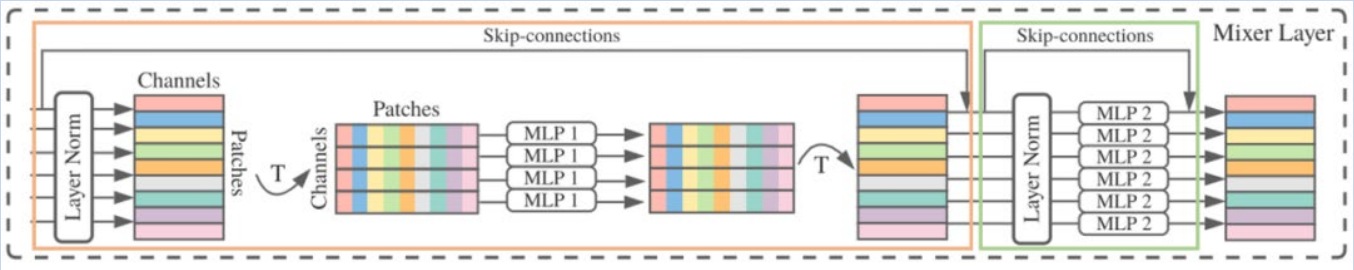
On the Connection between Local Attention and Dynamic Depth-wise Convolution

Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang

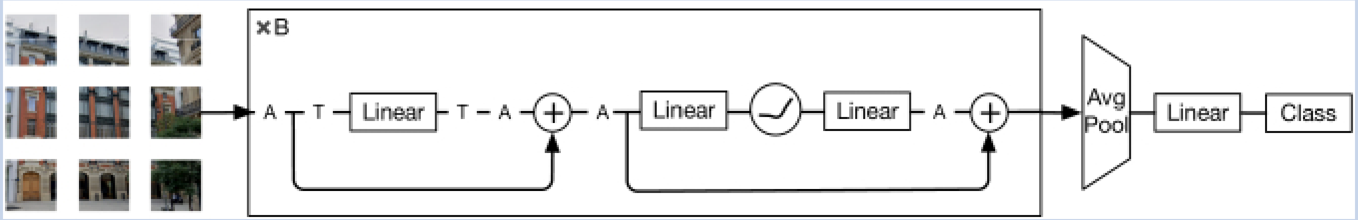
Vision Transformer: Attention



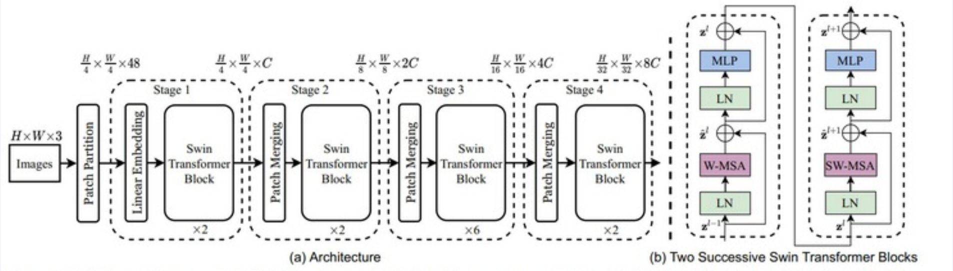
MLP-Mixer



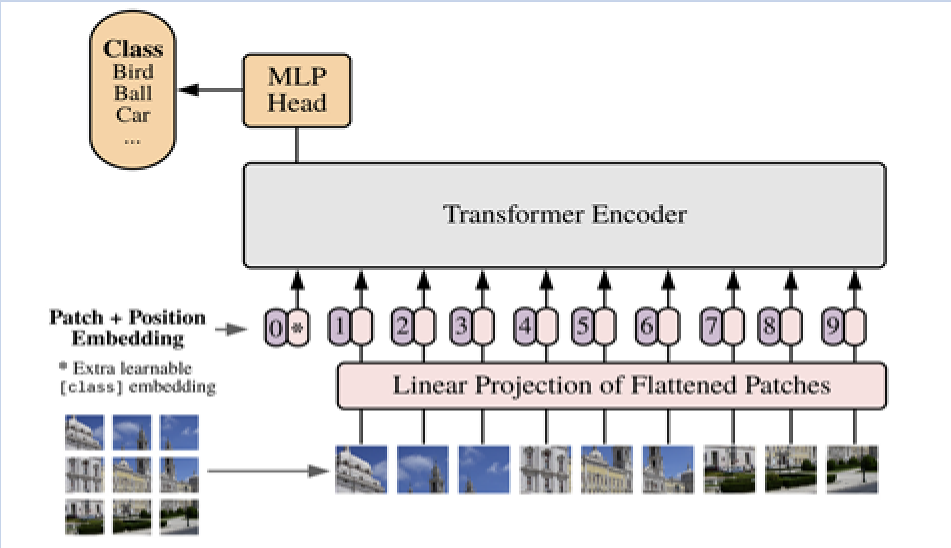
ResMLP



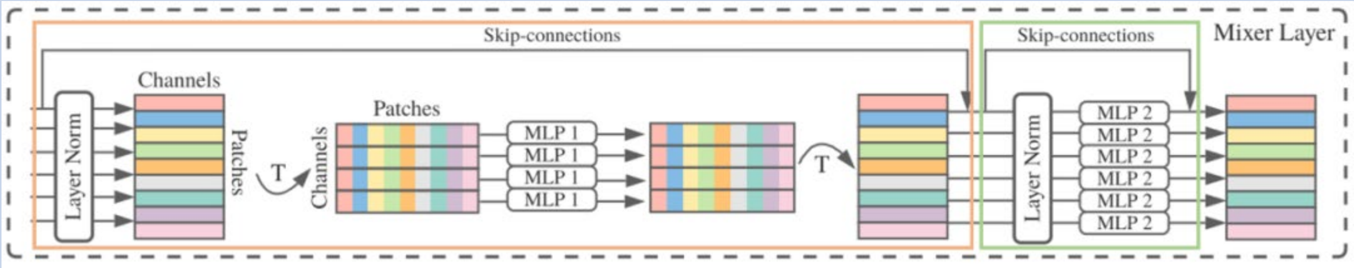
Swin: local attention



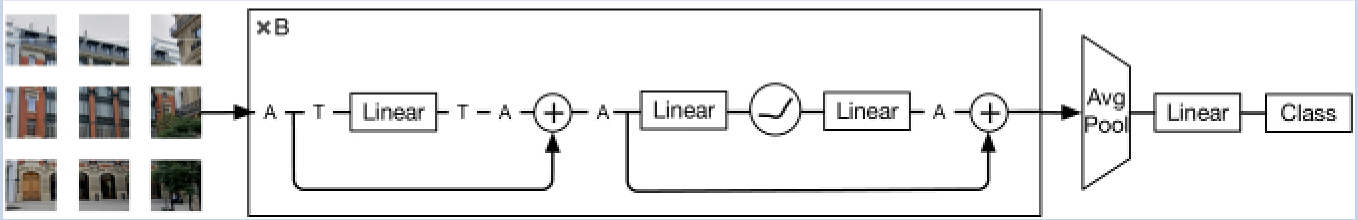
Vision Transformer: Attention



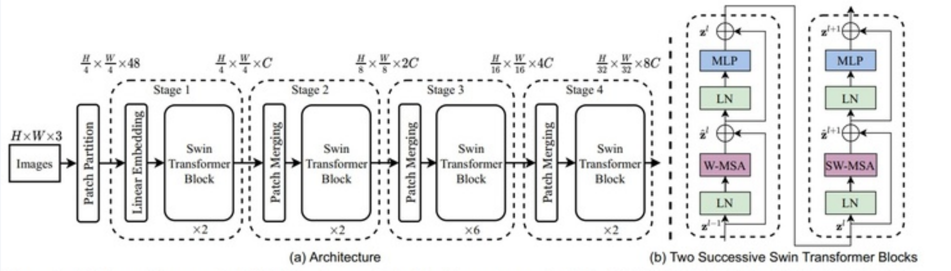
MLP-Mixer



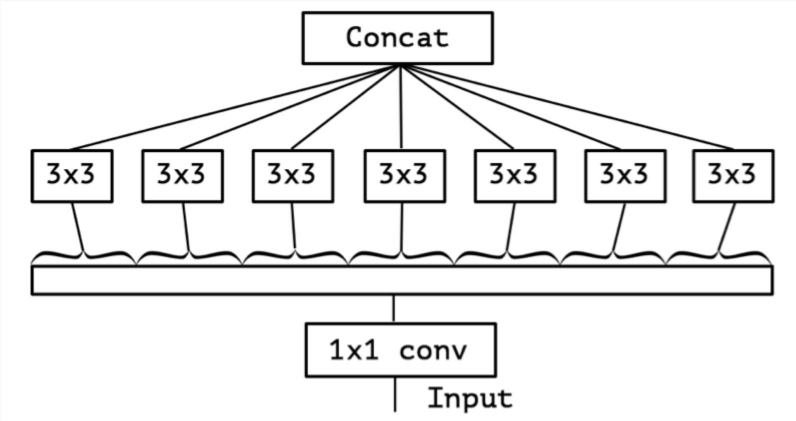
ResMLP



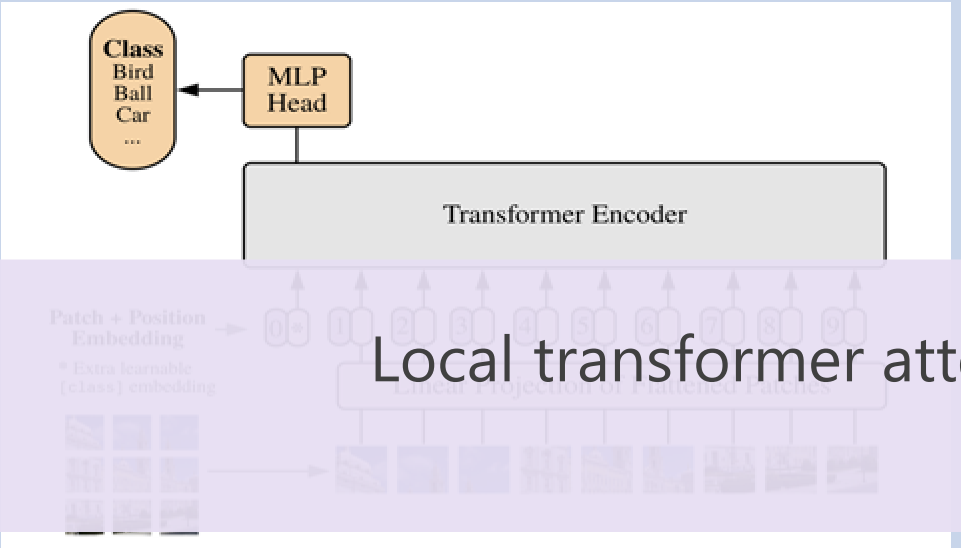
Swin: local attention



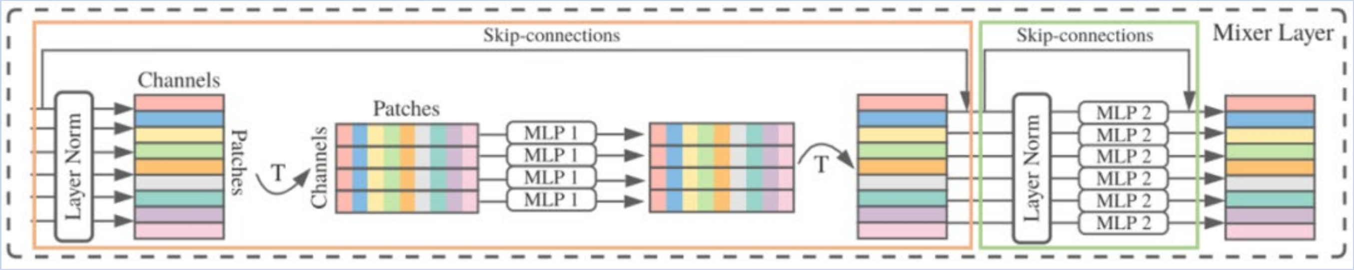
Depth-wise separable conv.



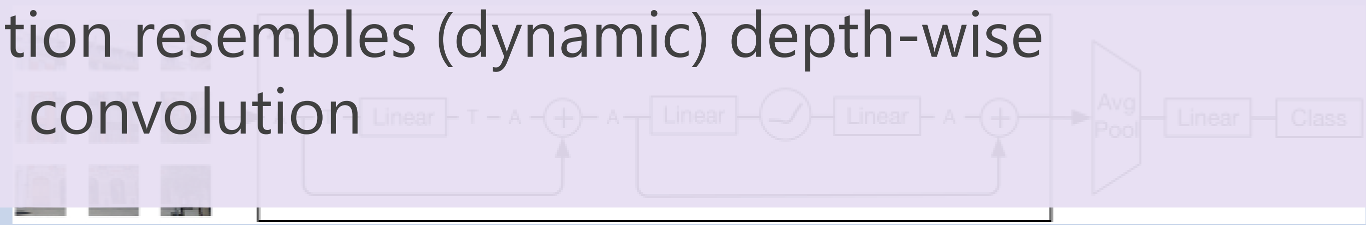
Vision Transformer: Attention



MLP-Mixer

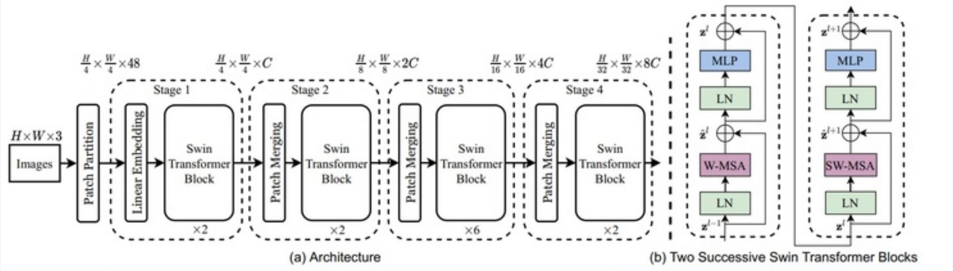


ResMLP

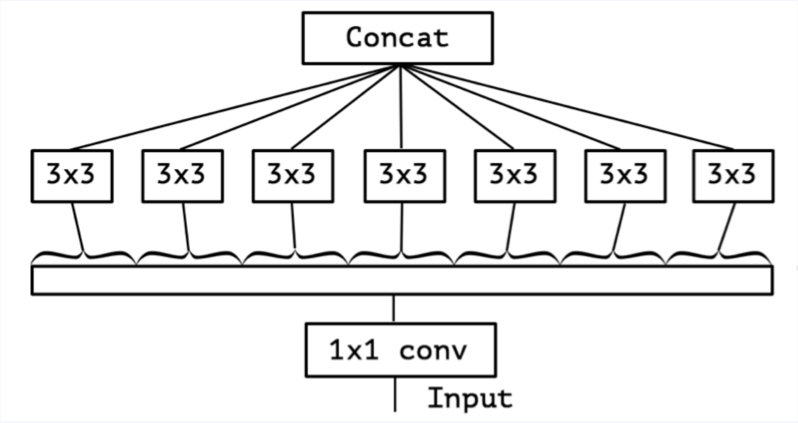


Local transformer attention resembles (dynamic) depth-wise convolution

Swin: local attention

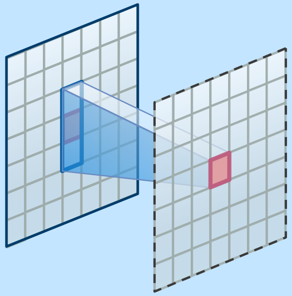


Depth-wise separable conv.



Local Attention vs Depth-wise Convolution: Local Connection

❑ Locally-connected



❑ Local attention

- Local aggregation

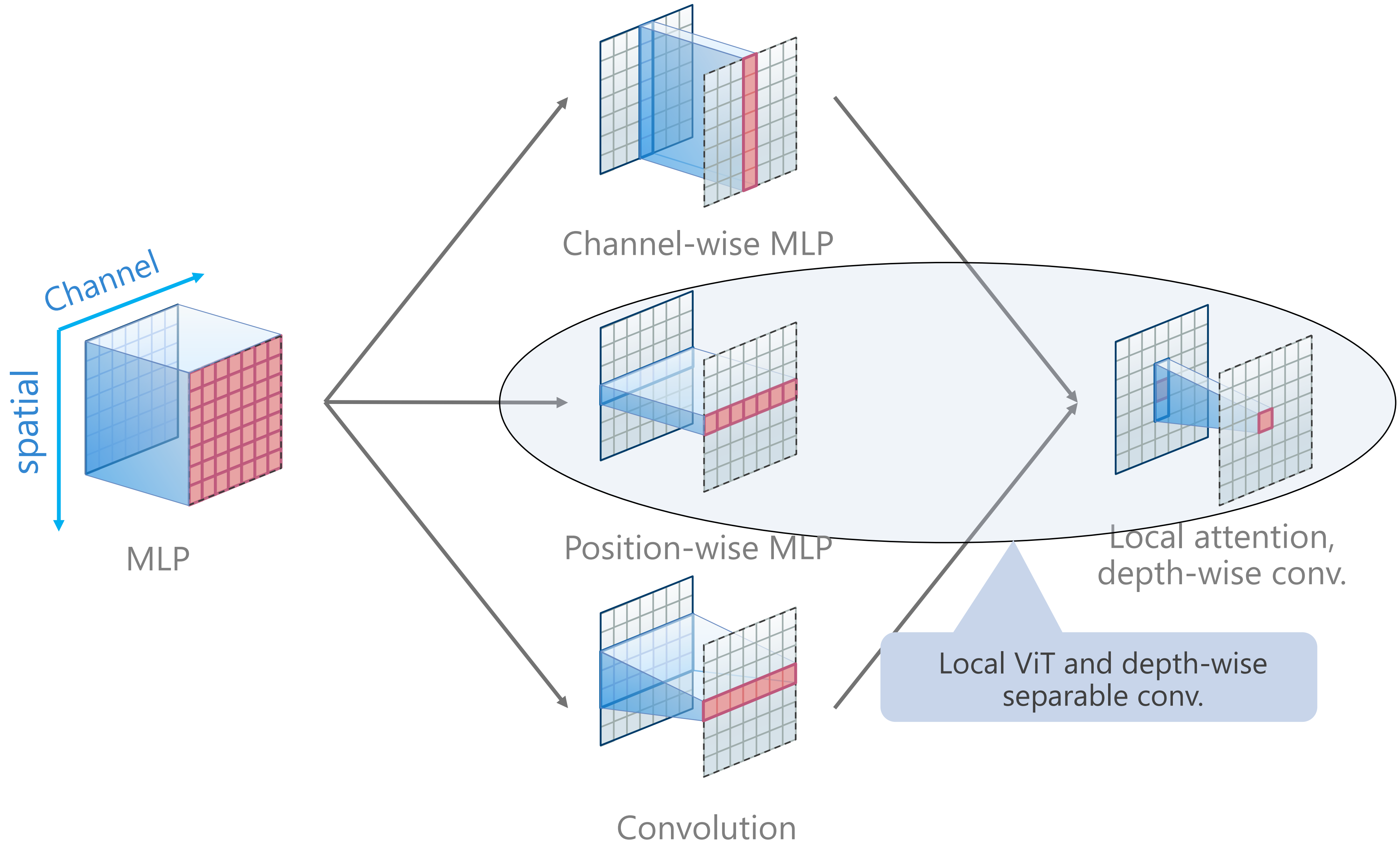
$$\mathbf{y}_i = \sum_{j=1}^{N_k} \mathbf{w}_{ij} \odot \mathbf{x}_{ij}$$

❑ Depth-wise convolution

- Local aggregation

$$\mathbf{y}_i = \sum_{j=1}^{N_k} \mathbf{w}_{\text{offset}(i,j)} \odot \mathbf{x}_{ij}$$

Local ViT and Depth-wise Sep. Conv.: Same Sparse Connectivity

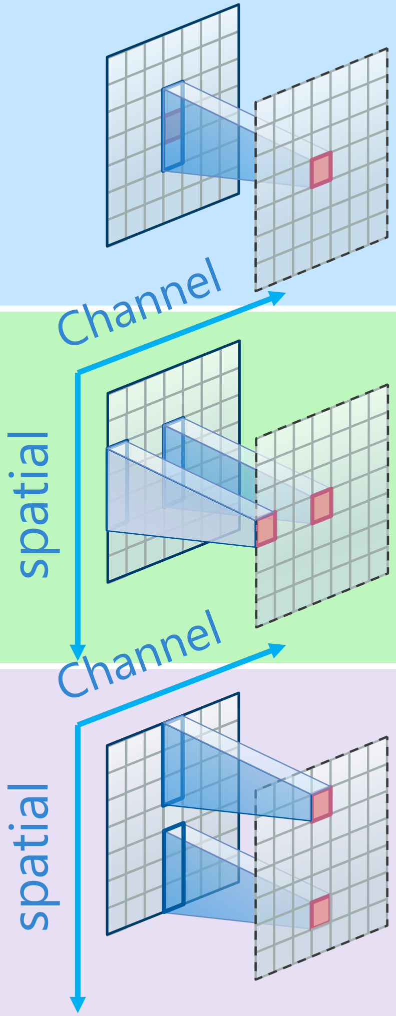


Local Attention vs Depth-wise Convolution: Weight Sharing

- ❑ Locally-connected

- ❑ Local attention
  - Weights shared across channels

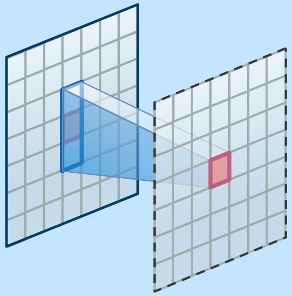
- ❑ Depth-wise convolution
  - Weights shared across positions



Local Attention vs (Dynamic) Depth-wise Convolution: Dynamic Weight

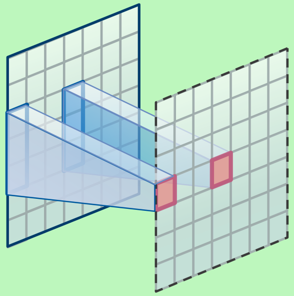
❑ Locally-connected

$$y = Wx$$



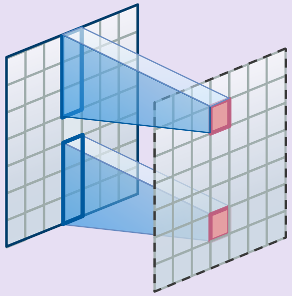
❑ Local attention

- Weights shared across channels
- Dynamic weight: $W_i = f(x)$



❑ Depth-wise convolution

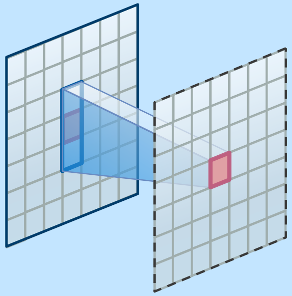
- Weights shared across positions
- Static weight: model parameter
- Can also be dynamic



Local Attention vs (Dynamic) Depth-wise Convolution: Dynamic Weight

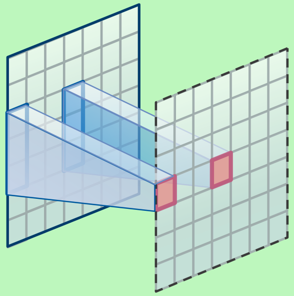
❑ Locally-connected

$$y = Wx$$



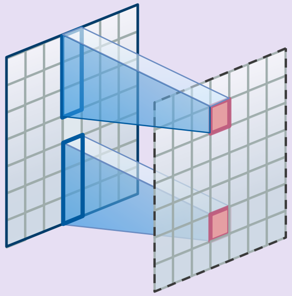
❑ Homogeneous Dynamic DW Conv.

- Weights shared across positions
- Dynamic weight: $W_i = f(\overline{x})$



❑ Inhomogeneous Dynamic DW Conv.

- Weights shared across channels
- Dynamic weight: $W_i = f(x_i)$



Local Attention vs DW-Conv.: Weight Sharing and Dynamic Weight

	Non-local sparse	Weight sharing across channels	Weight sharing across positions	Dynamic weight
Local attention	√	√		√
DW-Conv.	√		√	
Homogeneous dynamic DW-Conv.	√		√	√
Inhomogeneous dynamic DW-Conv.	√	√		√

Dynamic Depth-wise Convolution vs Attention

	ImageNet				COCO				ADE20K		
	#param.	FLOPs	top-1 acc.	real acc.	#param.	FLOPs	AP^{box}	AP^{mask}	#param.	FLOPs	mIoU
Swin-T	28M	4.5G	81.3	86.6	86M	747G	50.5	43.7	60M	947G	44.5
DW Conv.-T	24M	3.8G	81.3	86.8	82M	730G	49.9	43.4	56M	928G	45.5
D-DW Conv.-T	51M	3.8G	81.9	87.3	108M	730G	50.5	43.7	83M	928G	45.7
I-Dynamic-T	26M	3.95G	81.8	87.1	84M	741G	50.8	44.0	58M	939G	46.2
Swin-B	88M	15.4G	83.3	87.9	145M	986G	51.9	45.0	121M	1192G	48.1
DW Conv.-B	74M	12.9G	83.2	87.9	132M	924G	51.1	44.2	108M	1129G	48.3
D-DW Conv.-B	162M	13.0G	83.2	87.9	219M	924G	51.2	44.4	195M	1129G	48.0
I-Dynamic-B	80M	13.6G	83.4	88.0	137M	948G	51.8	44.8	114M	1153G	47.8

Dynamic Depth-wise Convolution vs Attention

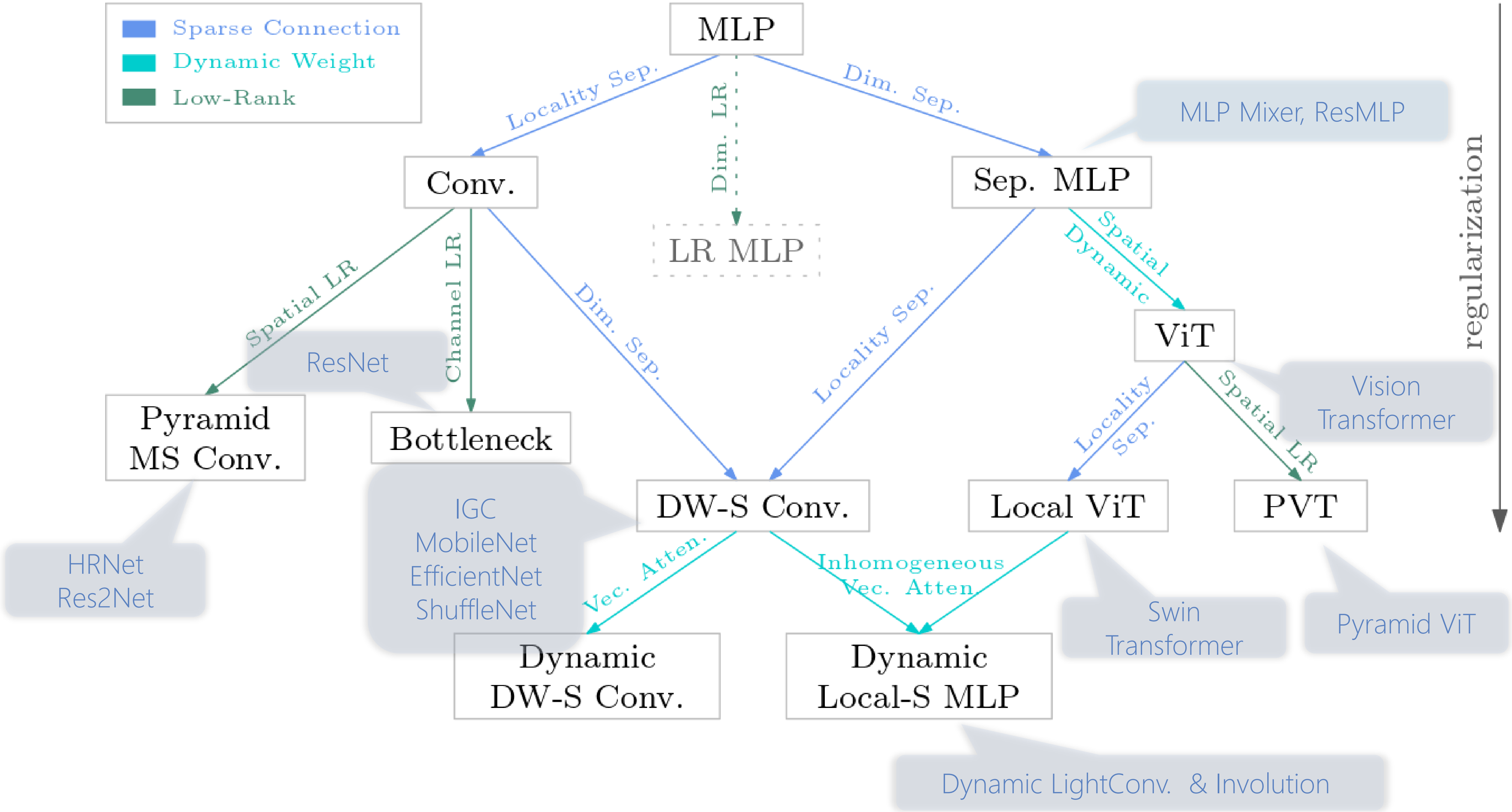
❑ Large scale pre-training

- Higher performance comes from the larger kernel size[1], eg 7x7. (Compared with traditional conv., such as 3x3)

	ImageNet1k fine-tuning			ADE20K fine-tuning		
	#param.	FLOPs	top-1 acc.	#param.	FLOPs	mIoU
Swin-B	88M	15.4G	85.2	121M	1192G	49.4
DW-Conv.-B	74M	12.9G	84.8	108M	1129G	50.1
D-DW-Conv.-B	162M	13.0G	85.0	195M	1129G	49.6

[1] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. Adv. Neural Inform. Process. Syst.

Relation Graph for Typical Networks



Codes are Available

https://github.com/Atten4Vis

Thanks!