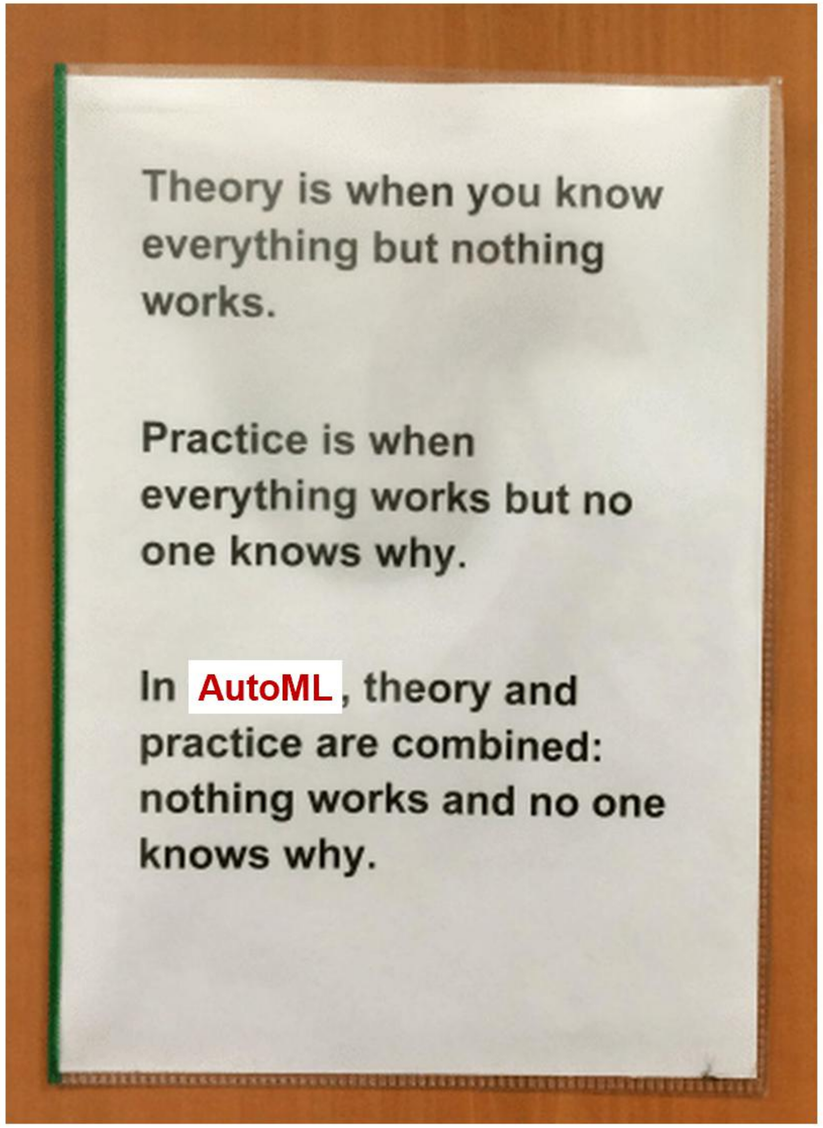
视觉基础模型架构设计新思路
——Inspiration from ViTs

张祥雨

旷视研究院

近年研究热点回顾

- 层数更深、性能更强的架构
- 轻量级架构、高推理效率
- 自动化模型设计、神经网络架构搜索（NAS）
- 动态模型
- Attention Models、Vision Transformers (ViTs)

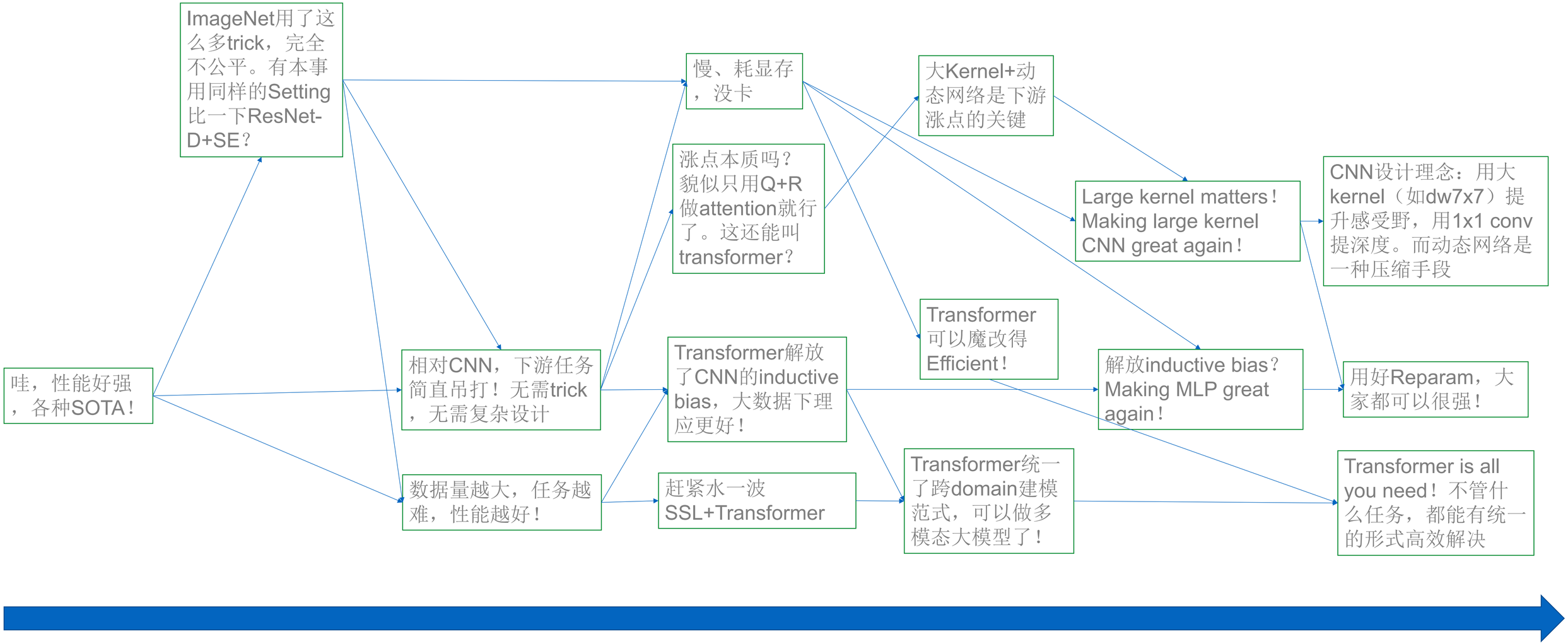


近年研究热点回顾

MEGVII 旷视

- 层数更深、性能更强的架构
- 轻量级架构、高推理效率
- 自动化模型设计、神经网络架构搜索（NAS）
- 动态模型
- Attention Models、**Vision Transformers (ViTs)**

如何看待Vision Transformers?



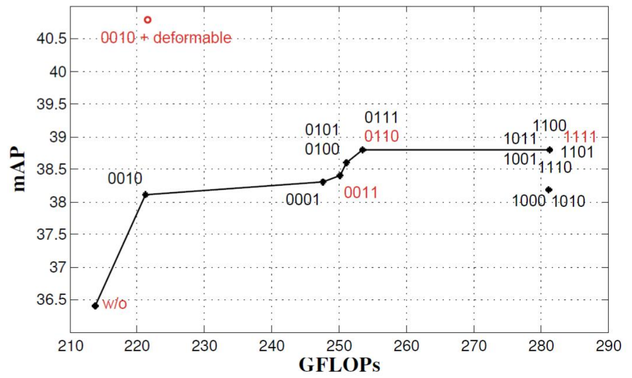
理解ViTs：潜在的优势

- 灵活的数据形式（Tensor，集合，序列，图）
- 长程关系建模能力
- 更强的表示能力 [1]
- 架构的合理性 [2]
- 对遮挡、噪声的稳健性 [3, 4]

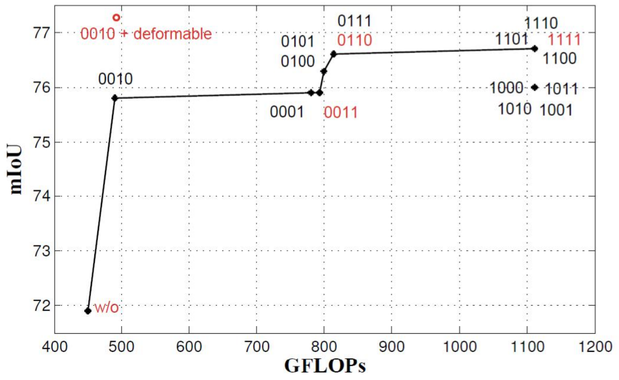
[1] Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. "On the relationship between self-attention and convolutional layers." ICLR 2020.
[2] Dong, Yihe, Jean-Baptiste Cordonnier, and Andreas Loukas. "Attention is not all you need: Pure attention loses rank doubly exponentially with depth.
[3] Xie, Cihang, et al. "Feature denoising for improving adversarial robustness." CVPR 2019.
[4] Paul, Sayak, and Pin-Yu Chen. "Vision transformers are robust learners."

Rethinking ViTs

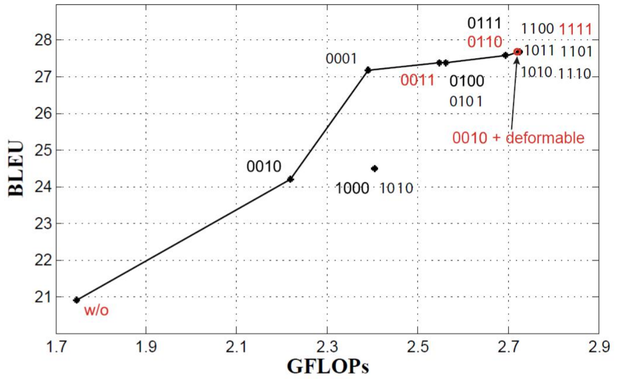
❖ Multi-head self-attention (MHSA)可能不是必须的



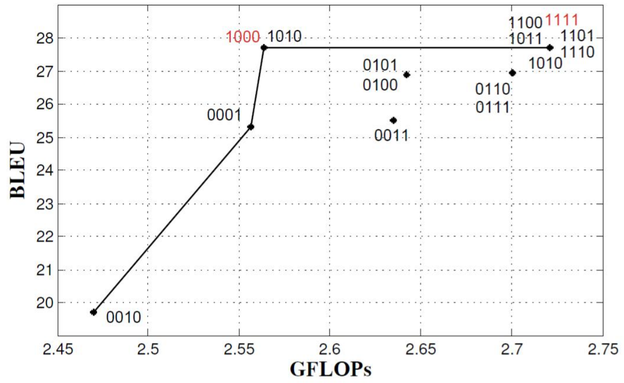
(a) Image Object detection on COCO



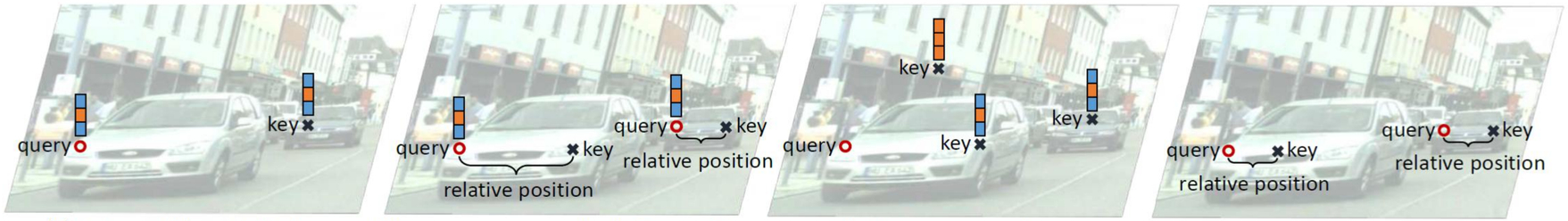
(b) Image Semantic Segmentation on Cityscapes



(c) Translation on newstest2014 (self-attention)



(d) Translation on newstest2014 (encoder-decoder attention)



(a) query and key content (b) query content and relative position (c) key content only (d) relative position only

Zhu, Xizhou, et al. "An empirical study of spatial attention mechanisms in deep networks." ICCV 2019.

Rethinking ViTs (cont’d)

❖ ViTs的设计要素

- ❖ Sparse connectivity
- ❖ Weight sharing
- ❖ Dynamic weight

Local attention: perform attention in local small windows						
Swin-T [37]	224^2	28M	4.5G	713.5	81.3	86.6
Swin-B [37]	224^2	88M	15.4G	263.0	83.3	87.9
Depth-wise convolution + point-wise 1×1 *convolution*						
DW-Conv.-T	224^2	24M	3.8G	928.7	81.3	86.8
DW-Conv.-B	224^2	74M	12.9G	327.6	83.2	87.9
D-DW-Conv.-T	224^2	51M	3.8G	897.0	81.9	87.3
D-DW-Conv.-B	224^2	162M	13.0G	322.4	83.2	87.9

Table 3: Comparison results on COCO object detection and ADE semantic segmentation.

	COCO Object Detection						ADE20K Semantic Segmentation		
	#param.	FLOPs	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	#param.	FLOPs	mIoU
Swin-T	86M	747G	50.5	69.3	54.9	43.7	60M	947G	44.5
DW Conv.-T	82M	730G	49.9	68.6	54.3	43.4	56M	928G	45.5
D-DW Conv.-T	108M	730G	50.5	69.5	54.6	43.7	83M	928G	45.7
Swin-B	145M	986G	51.9	70.9	56.5	45.0	121M	1192G	48.1
DW Conv.-B	132M	924G	51.1	69.6	55.4	44.2	108M	1129G	48.3
D-DW Conv.-B	219M	924G	51.2	70.0	55.4	44.4	195M	1129G	48.0

[1] Han, Qi, et al. "Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight.“
[2] Zhao, Yucheng, et al. "A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP."
Rao, Yongming, et al. "Global filter networks for image classification."

ViTs到底做对了什么?

MEGVII 旷视

- ViTs的设计要素、潜在优势等**并非独有**

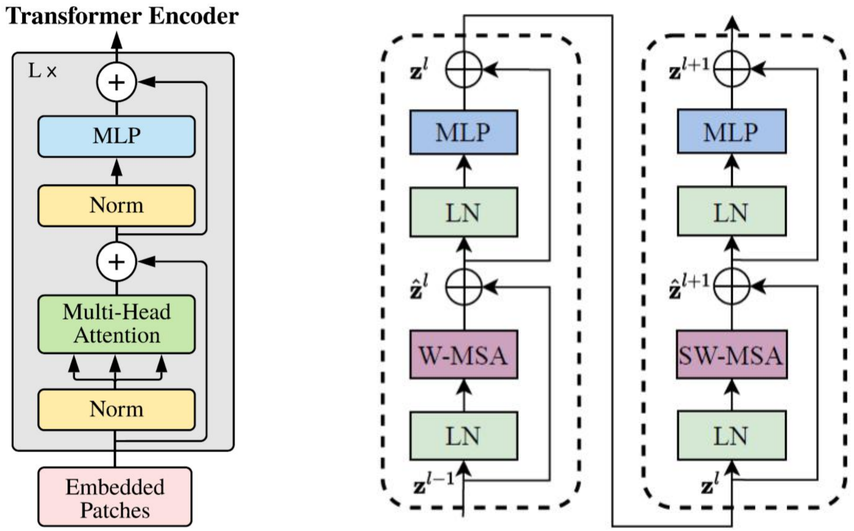
- What can we learn from vision transformers?
  - Large kernel design
  - High-order relation modeling

Inspiration from Vision Transformers

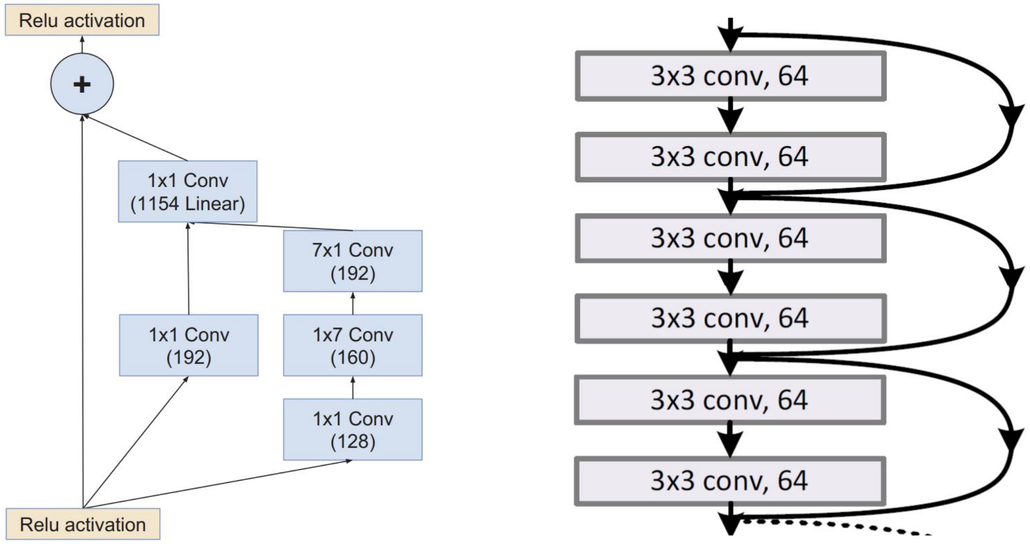
- **Large kernel models**
- High-order relation modeling

Spatial Modeling

- Vision Transformers
  - Global MHSA [1]
  - Local MHSA (e.g. >=7x7) [2, 3]



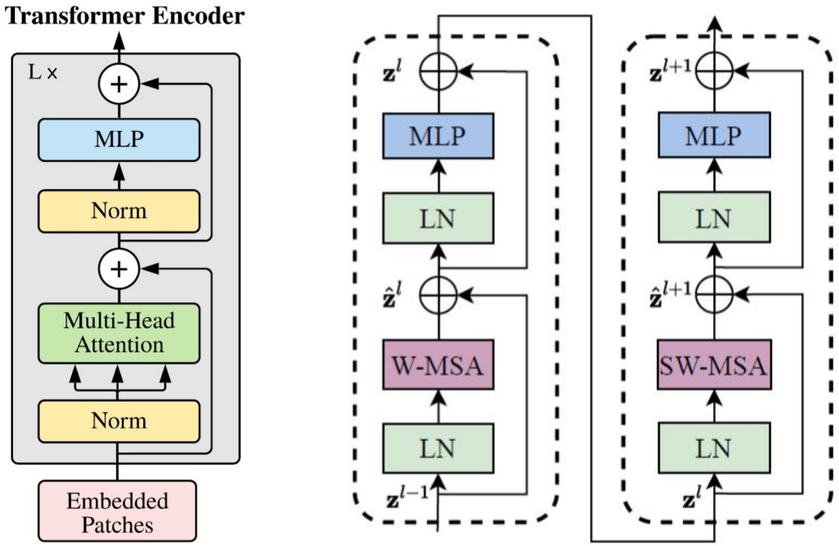
- CNNs
  - Large kernels (e.g. >= 5x5) [4]
  - Stack of 3x3 (DW, Group) Convs [5]

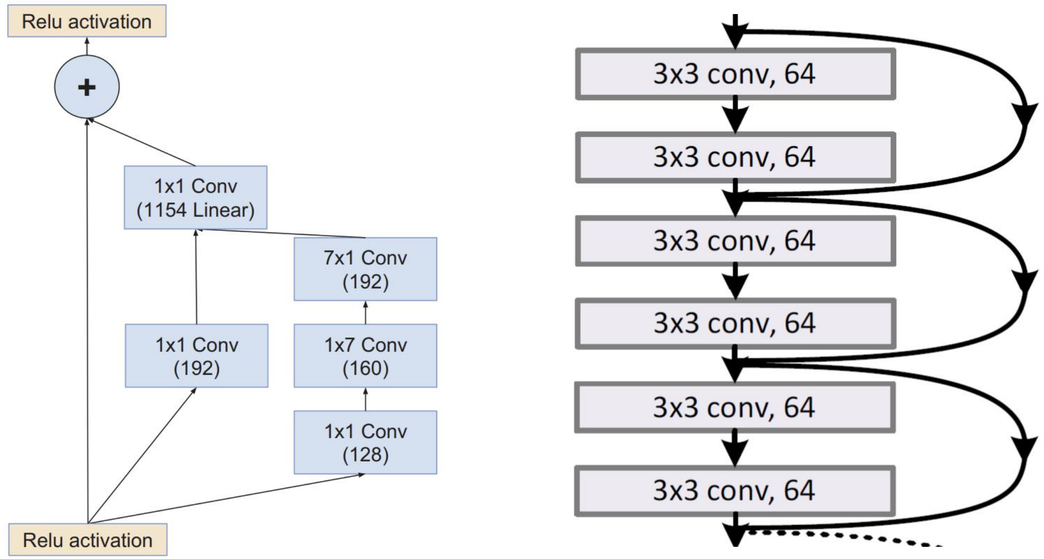


[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." ICLR 2021.
[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021.
[3] Dong, Xiaoyi, et al. "Cswin transformer: A general vision transformer backbone with cross-shaped windows."
[4] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI 2017.
[5] He et al. "Deep residual learning for image recognition." CVPR 2016.

Spatial Modeling

- Vision Transformers
  - Global MHSA [1]
  - Local MHSA (e.g. >=7x7) [2]
- CNNs
  - Large kernels (e.g. >= 5x5) [3]
  - **Stack of 3x3 (DW, Group) Convs [4]**





[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." ICLR 2021.
[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021.
[3] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI 2017.
[4] He et al. "Deep residual learning for image recognition." CVPR 2016.

Advantages of Large Kernels

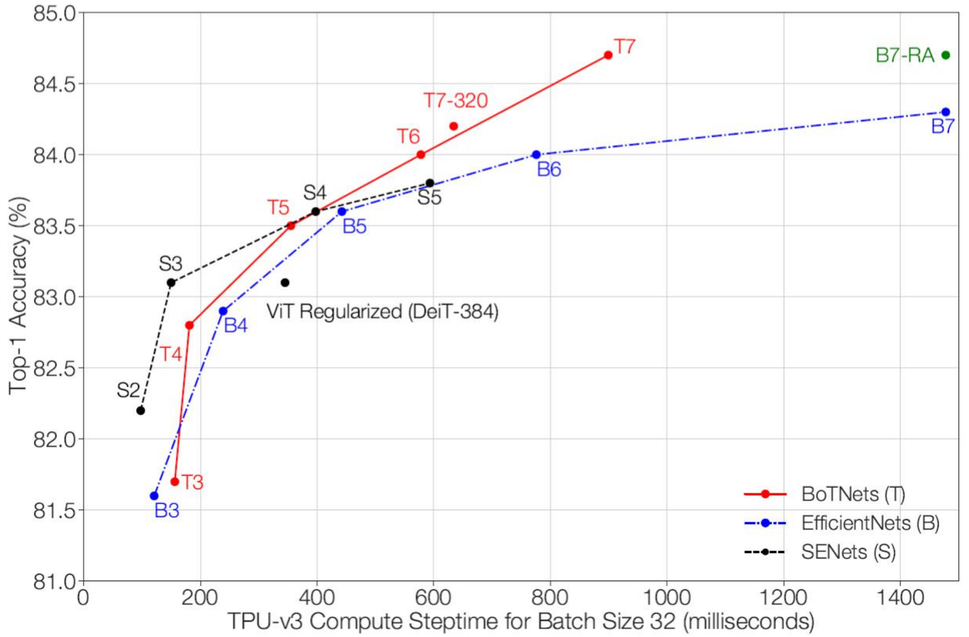
❖大卷积核可以更高效地提升有效感受野（ERF）

$$\sqrt{\mathrm{Var}[S_n]} = \sqrt{n}\sqrt{\sum_{m=0}^{k-1}\frac{m^2}{k} - \left(\sum_{m=0}^{k-1}\frac{m}{k}\right)^2} = \sqrt{\frac{n(k^2-1)}{12}} = O(k\sqrt{n})$$

Luo, Wenjie, et al. "Understanding the effective receptive field in deep convolutional neural networks." NIPS 2016.

Advantages of Large Kernels (cont’d)

- 大卷积核可以部分回避模型深度增加带来的优化难题
  - VGG-style models 难以做深 [1]
  - ResNets 的有效深度可能很浅 [2, 3]



[1] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." ICLR 2015.

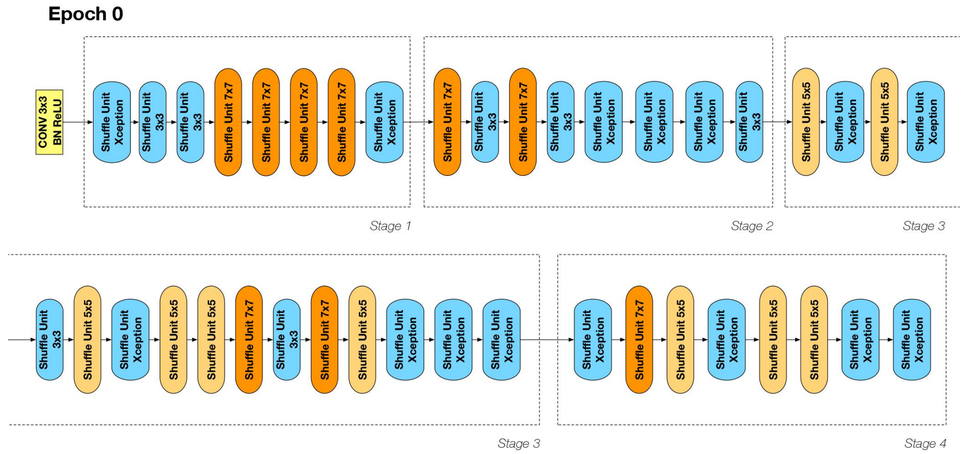
[2] Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks." NIPS 2016.

[3] De, Soham, and Samuel L. Smith. "Batch normalization biases residual blocks towards the identity function in deep networks." NeurIPS 2020.

Advantages of Large Kernels (cont’d)

- 大卷积核对FCN-based的下游任务提升明显
  - Detection (e.g. deformable conv [1], DetNAS [2])
  - Segmentation (e.g. global conv [3], dilated conv [4])

k	3	5	7	9	11
Score (GCN)	70.1	71.1	72.8	73.4	73.7
Score (Stack)	69.8	71.8	71.3	69.5	67.5



[1] Dai, Jifeng, et al. "Deformable convolutional networks." ICCV 2017.
[2] Chen, Yukang, et al. "Detnas: Backbone search for object detection." NeurIPS 2019.
[3] Peng, Chao, et al. "Large kernel matters--improve semantic segmentation by global convolutional network." CVPR 2017.
[4] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." TPAMI 2017.

为什么大卷积核CNN没有成为主流?

MEGVII 旷视

- 架构超参 [1, 2]
  - 宽度
  - 深度
  - 输入分辨率
  - **卷积核大小**

[1] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." ICML 2019.
[2] Radosavovic, Ilija, et al. "Designing network design spaces." CVPR 2020.

Large Kernel CNN Design

MEGVII 旷视

- 问题：ImageNet的局限性
  - ImageNet分类可能更偏向纹理特征 [1]
  - ImageNet分类任务对感受野要求不高 [2, 3]

Pretrained Model	ResNet50	ResNet50-GCN
ImageNet cls err (%)	7.7	7.9
Seg. Score (Baseline)	65.7	71.2
Seg. Score (GCN + BR)	72.3	**72.5**

- 对策：更强的训练和数据增广策略

Backbone	epochs	AP^{bb}	AP^{mk}
R50	12	39.0	35.0
BoT50	12	39.4 (+ **0.4**)	35.3 (+ **0.3**)
R50	24	41.2	36.9
BoT50	24	42.8 (+ **1.6**)	38.0 (+ **1.1**)
R50	36	42.1	37.7
BoT50	36	43.6 (+ **1.5**)	38.9 (+ **1.2**)
R50	72	42.8	37.9
BoT50	72	43.7 (+ **0.9**)	38.7 (+ **0.8**)

[1] Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." ICLR 2019.
[2] Peng, Chao, et al. "Large kernel matters--improve semantic segmentation by global convolutional network." CVPR 2017.
[3] Srinivas, Aravind, et al. "Bottleneck transformers for visual recognition." CVPR 2021.

Large Kernel CNN Design (Cont'd)

MEGVII 旷视

- 问题：大卷积核不够高效
- 对策
  - 更浅的结构 [1]
  - 卷积核分解 [2]
  - FFT Conv [3]
  - 稀疏算子 [4]

	downsp. rate (output size)	Swin-T
stage 1	4× (56×56)	concat 4×4, 96-d, LN
		[win. sz. 7×7, dim 96, head 3] × 2
stage 2	8× (28×28)	concat 2×2, 192-d , LN
		[win. sz. 7×7, dim 192, head 6] × 2
stage 3	16× (14×14)	concat 2×2, 384-d , LN
		[win. sz. 7×7, dim 384, head 12] × 6
stage 4	32× (7×7)	concat 2×2, 768-d , LN
		[win. sz. 7×7, dim 768, head 24] × 2

(a) Various frameworks

Method	Backbone	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	#param.	FLOPs	FPS
Cascade Mask R-CNN	R-50	46.3	64.3	50.5	82M	739G	18.0
	Swin-T	**50.5**	**69.3**	**54.9**	86M	745G	15.3
ATSS	R-50	43.5	61.9	47.0	32M	205G	28.3
	Swin-T	**47.2**	**66.5**	**51.3**	36M	215G	22.3
RepPointsV2	R-50	46.5	64.6	50.3	42M	274G	13.6
	Swin-T	**50.0**	**68.5**	**54.2**	45M	283G	12.0
Sparse R-CNN	R-50	44.5	63.4	48.2	106M	166G	21.0
	Swin-T	**47.9**	**67.3**	**52.3**	110M	172G	18.4

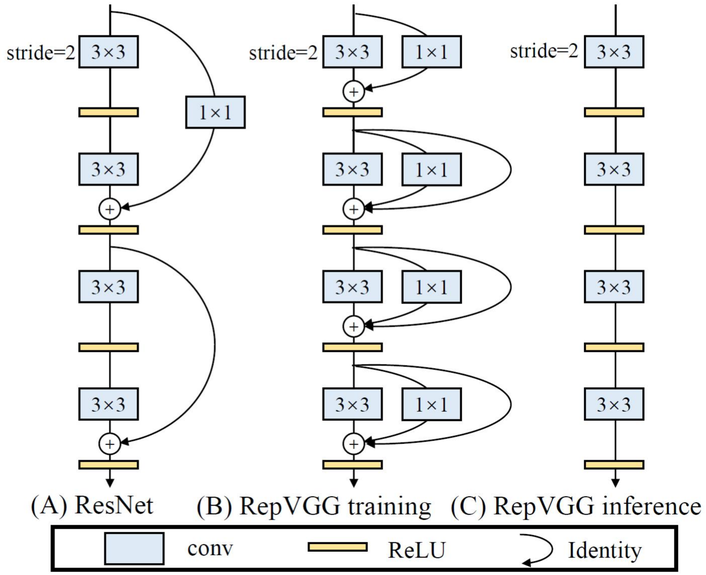
Table 3: Comparison results on COCO object detection and ADE semantic segmentation.

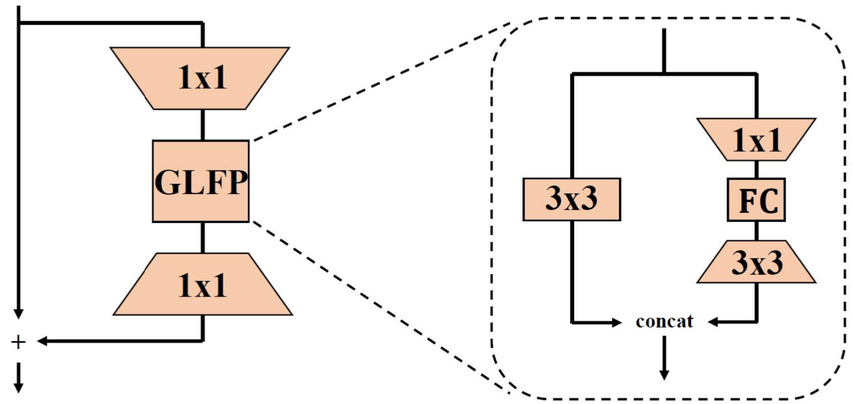
	COCO Object Detection						ADE20K Semantic Segmentation		
	#param.	FLOPs	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	#param.	FLOPs	mIoU
Swin-T	86M	747G	50.5	69.3	54.9	43.7	60M	947G	44.5
DW Conv.-T	82M	730G	49.9	68.6	54.3	43.4	56M	928G	45.5
D-DW Conv.-T	108M	730G	50.5	69.5	54.6	43.7	83M	928G	45.7
Swin-B	145M	986G	51.9	70.9	56.5	45.0	121M	1192G	48.1
DW Conv.-B	132M	924G	51.1	69.6	55.4	44.2	108M	1129G	48.3
D-DW Conv.-B	219M	924G	51.2	70.0	55.4	44.4	195M	1129G	48.0

[1] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021.
[2] Han, Qi, et al. "Demystifying Local Vision Transformer: Sparse Connectivity, Weight Sharing, and Dynamic Weight."
[3] Rao, Yongming, et al. "Global filter networks for image classification."
[4] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." ICLR 2021.

Large Kernel CNN Design (Cont’d)

- 问题：大卷积核难以兼顾局部特征
- 对策
  - 结构重参数化方法 [1, 2]





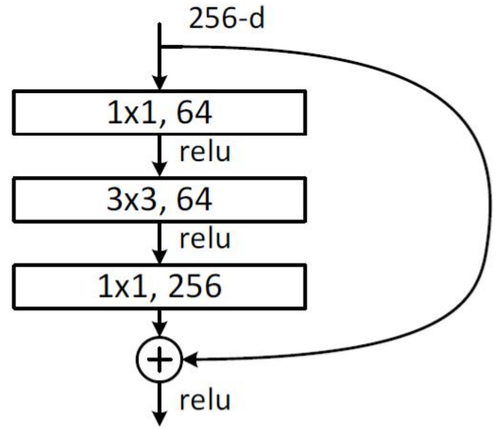
[1] Ding, Xiaohan, et al. "Repvgg: Making vgg-style convnets great again." CVPR 2021.
[2] Ding, Xiaohan, et al. "RepMLP: Re-parameterizing Convolutions into Fully-connected Layers for Image Recognition."

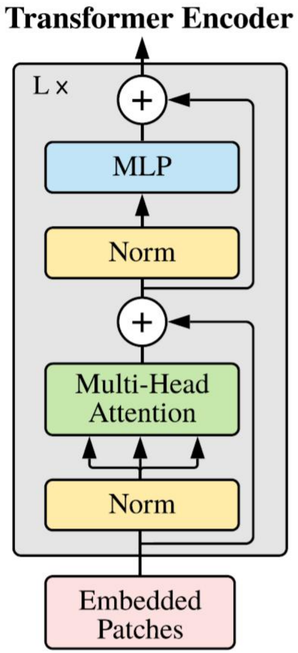
Large Kernel CNN Design (Cont’d)

❖ 问题：spatial modeling和semantic modeling对深度的要求并不一致

- ❖ Spatial modeling：感受野的大小、Relation的阶数
- ❖ Semantic modeling：语义的复杂程度

❖ 对策：摒弃堆叠单一重复单元的架构设计范式，对Spatial和Depth分别设计





小结：通向大感受野视觉模型设计

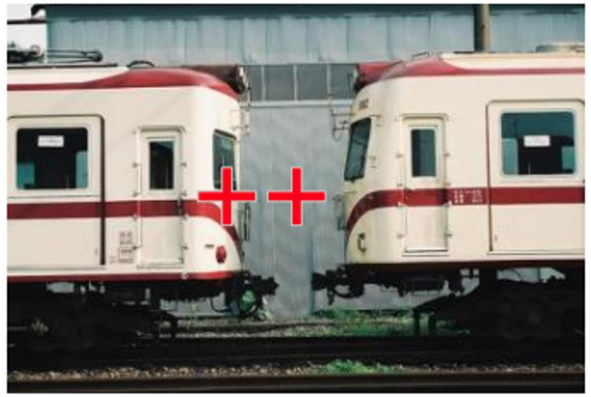
- 采用更强的训练和数据增广策略
- 浅且高效的Spatial算子
- 使用结构重参数化等方法添加架构先验
- 对Spatial和Depth分别设计

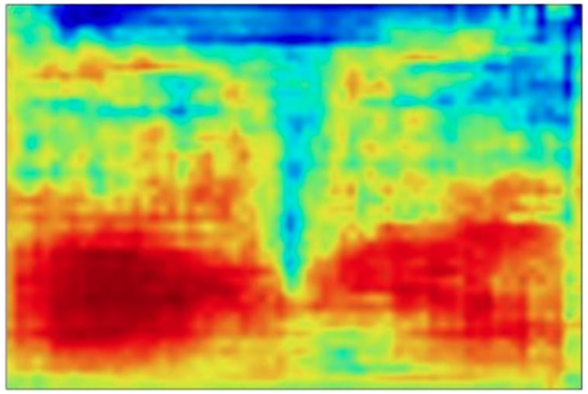
Inspiration from Vision Transformers

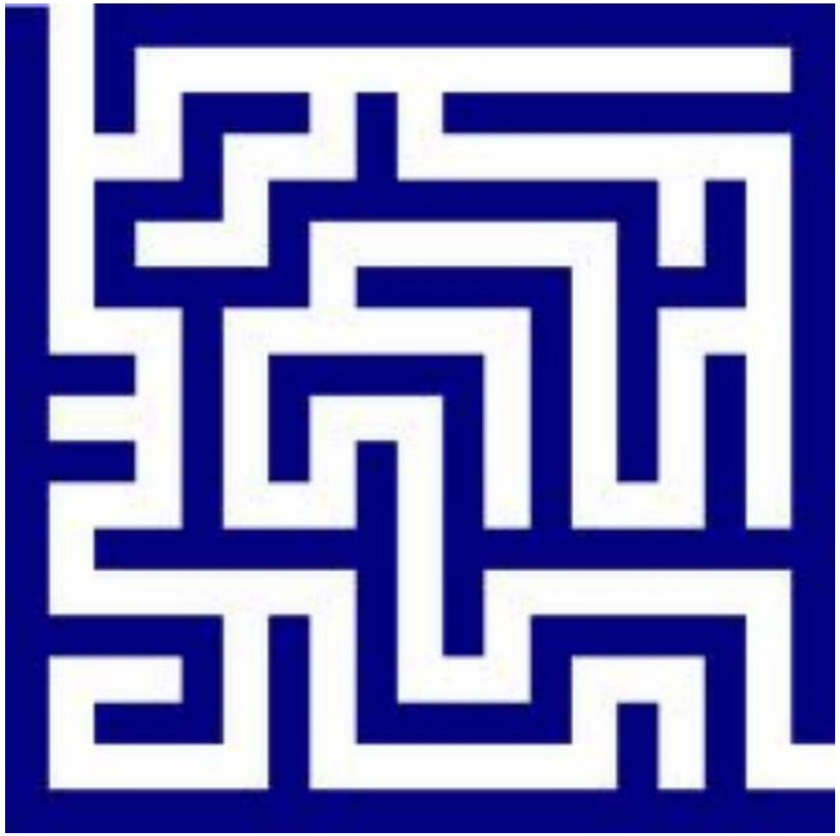
- Large kernel models
- **High-order relation modeling**

High-order Relation Modeling

❖MHSA擅长建模长程依赖关系。但是对于很多视觉问题，我们还需要处理高阶依赖关系

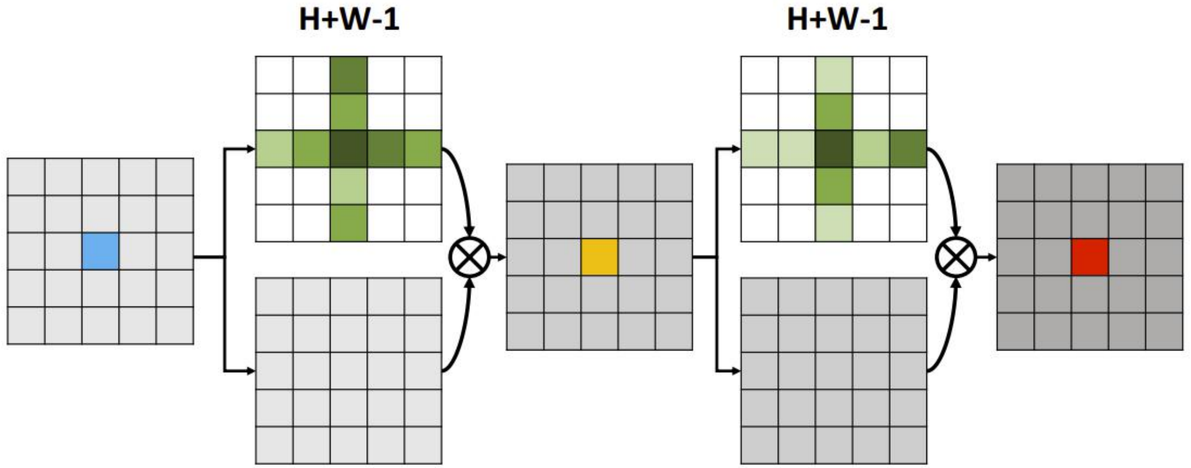






High-order Relation in MHSA

- 堆叠多层MHSA [1]
- High-order attention
- Recurrent attention
- 缺点：较为低效



[1] Huang, Zilong, et al. "Ccnet: Criss-cross attention for semantic segmentation." ICCV 2019.

Learnable Tree Filters (LTFs)

❖ Non-local attention [1]

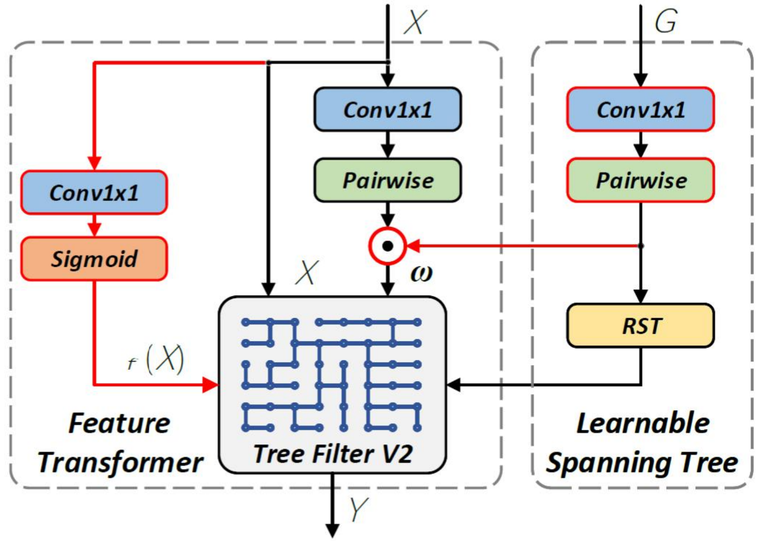
$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j)$$

❖ LTF v1 [2]

$$y_i = \frac{1}{z_i} \sum_{\forall j \in \mathcal{V}} S_{\mathcal{G}_T}(E_{j,i})\, x_j, \quad \text{where } S_{\mathcal{G}_T}(E_{j,i}) = \exp(- \sum_{\forall (k,m) \in E_{j,i}} \omega_{k,m})$$

❖ LTF v2 [3]

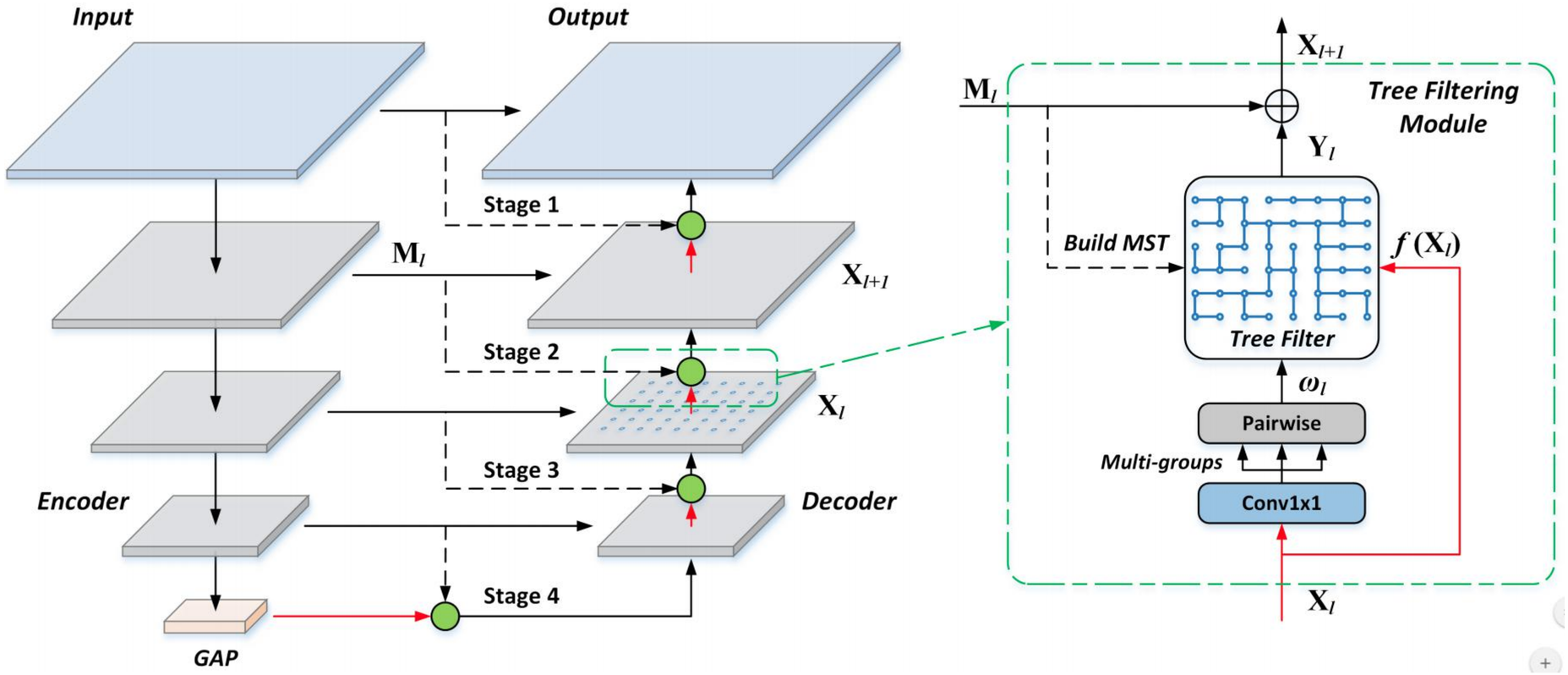
$$y_i = \frac{1}{z_i} \sum_{\forall x_j \in X} S_{\mathcal{G}_T}(E_{j,i}) \exp(-\beta)^{|E_{j,i}|} f(x_j) x_j$$



[1] Wang, Xiaolong, et al. "Non-local neural networks." CVPR 2018.
[2] Song, Lin, et al. "Learnable tree filter for structure-preserving feature transform." NeurIPS 2019.
[3] Song, Lin, et al. "Rethinking learnable tree filter for generic feature transform." NeurIPS 2020.

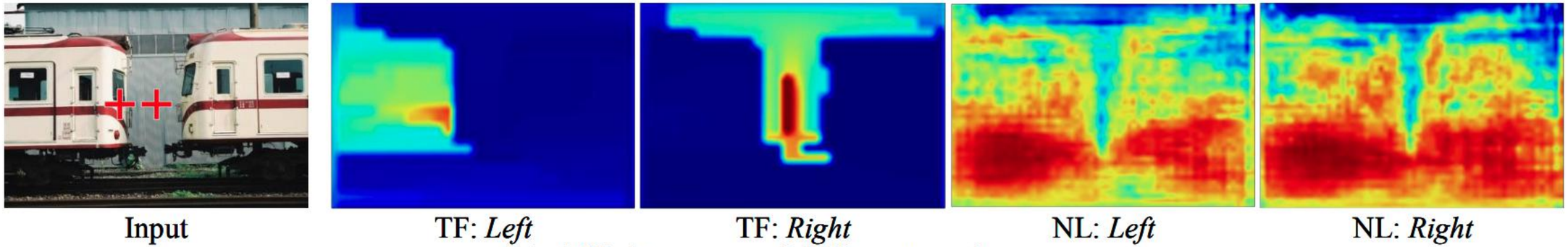
Advantages of LTFs

❖即插即用的模块，可以自由替换已有的attention module

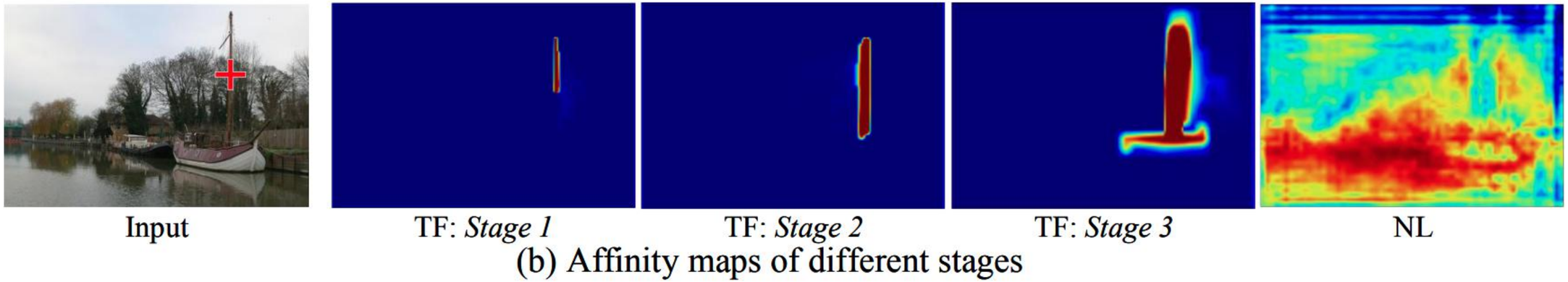


Advantages of LTFs (cont’d)

❖对高阶长程依赖的建模能力



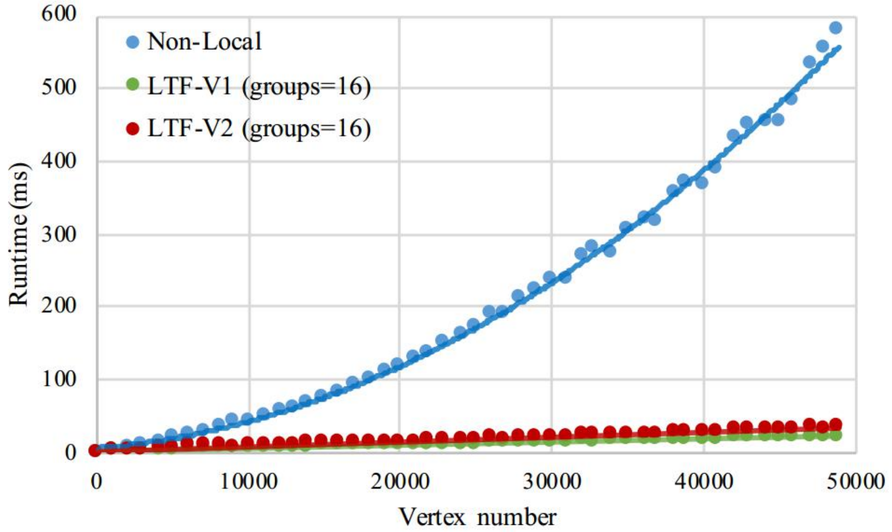
(a) Affinity maps of different anchors



(b) Affinity maps of different stages

Advantages of LTFs (cont’d)

❖更强的性能，更低的开销



Model	Stage	AP_{box}	AP^{50}_{box}	AP^{75}_{box}	AP_{seg}	AP^{50}_{seg}	AP^{75}_{seg}	#FLOPs	#Params
ResNet-50 (1x)	-	38.8	58.7	42.4	35.2	55.6	37.6	279.4B	44.4M
+Non-Local [11]	4	39.5	59.6	42.7	35.6	56.7	37.6	+10.67B	+2.09M
+CCNet [12]	345	40.1	60.4	44.1	36.0	57.4	38.4	+16.62B	+6.88M
+LatentGNN [14]	345	40.6	61.3	44.5	36.6	58.1	39.2	+3.59B	+1.07M
+GCNet [15]	All	40.7	61.0	44.2	36.7	58.1	39.2	+0.35B	+10.0M
+LTF-V1	345	40.0	60.4	43.7	36.1	57.5	38.4	+0.31B	+0.06M
+LTF-V2	3	40.1	59.9	43.9	36.0	57.1	38.3	+0.43B	+0.02M
	4	40.6	61.0	44.4	36.6	58.2	39.0	+0.26B	+0.04M
	5	40.2	60.5	43.6	36.1	57.5	38.4	+0.17B	+0.08M
	345	**41.2**	**61.6**	**45.2**	**37.0**	**58.4**	**39.5**	+0.68B	+0.14M

MEGVII 旷视

Q & A