Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation

Jiaqi Gu[1], Hyoukjun Kwon[2], Dilin Wang[2], Wei Ye[2], Meng Li[2], Yu-Hsin Chen[2], Liangzhen Lai[2], Vikas Chandra[2], David Z. Pan[1]

[1]University of Texas at Austin

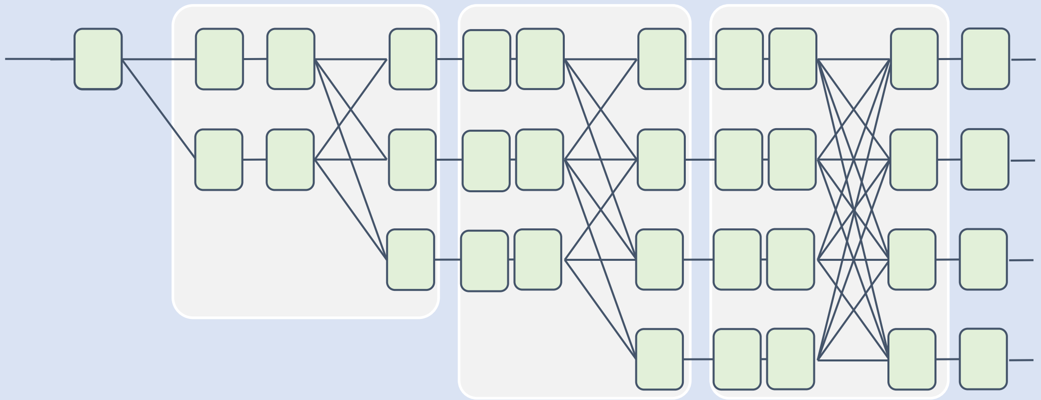
[2]Meta Platforms Inc.

CVPR 2022

Computer Vision Workloads

- Core applications
  - Not just classification..
  - Dense prediction vision tasks
    - **Semantic segmentation**
    - Object detection
    - Pose estimation
    - …
- Performance-Efficiency trade-off
  - **Low hardware cost** on edge devices
  - **High-performance** on dense prediction tasks

Evolve from CNN to ViTs



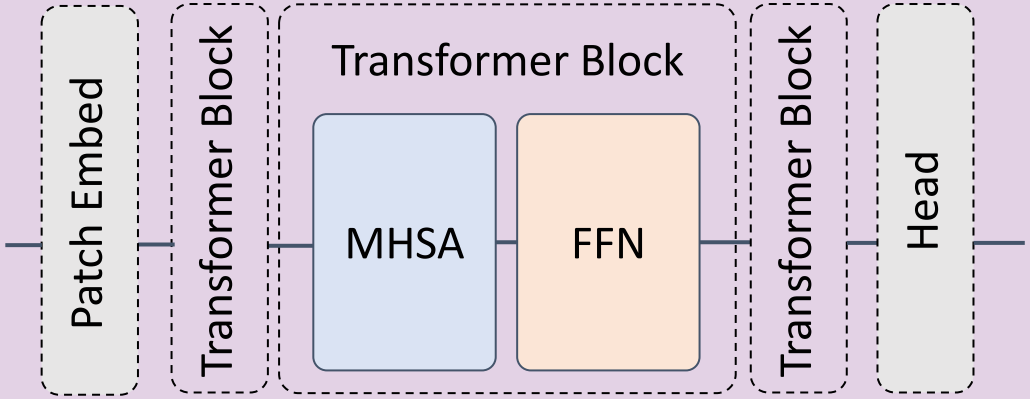
HRNet [1]

☺ **Multi-scale, Cross-resolution: *HR***

☹ Limited receptive field: *Conv*

☹ High complexity: *multi-branch*

[1] Image source: J. Wang, et al., “Deep high-resolution representation learning for visual recognition,” TPAMI, 2019.



ViT [2]

☹ Single-scale, Low-res.: *Sequential*

☺ **Large receptive field: *attention***

☹ High complexity: *attention*

[2] Image source: A. Dosovitskiy, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.

inherit

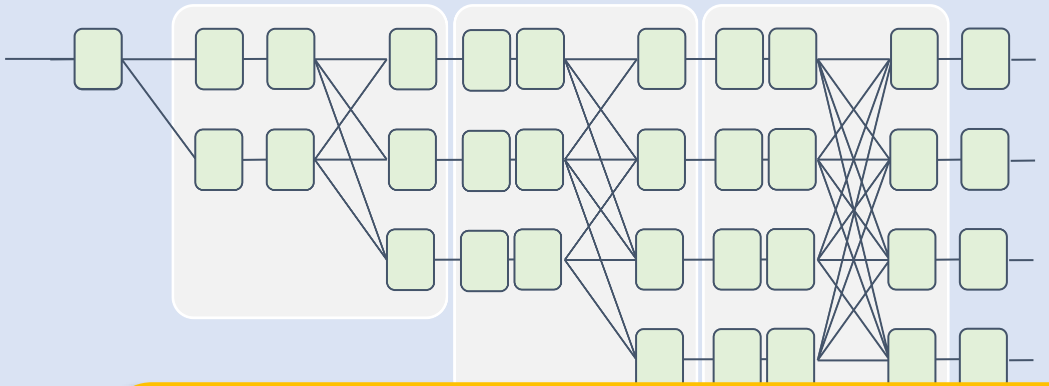
Multi-scale, Cross-resolution .... HR/multi-branch

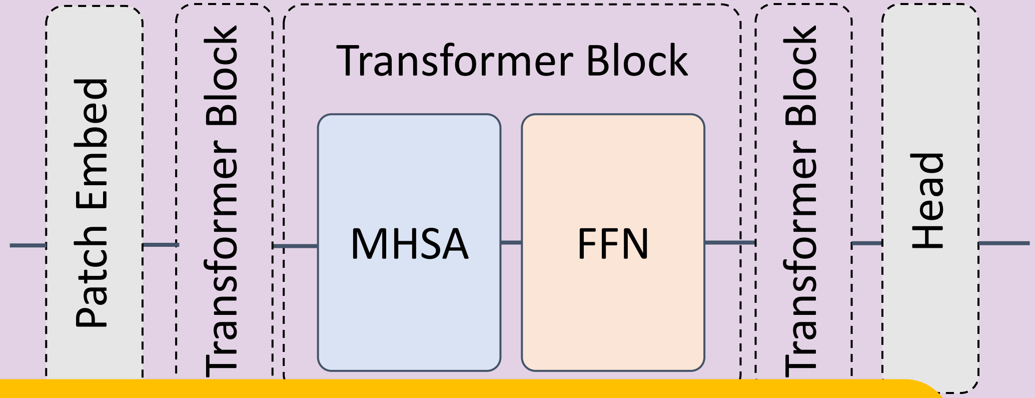
Large receptive field .... Self-attention

evolve

High efficiency .... **?**

Evolve from CNN to ViTs





Need **Synergistic Customization**

☺ M

☹

☹ High complexity: *multi-branch*

[1] Image source: J. Wang, et al., “Deep high-resolution representation learning for visual recognition,” TPAMI, 2019.

ial

n

☹ High complexity: *attention*

[2] Image source: A. Dosovitskiy, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.

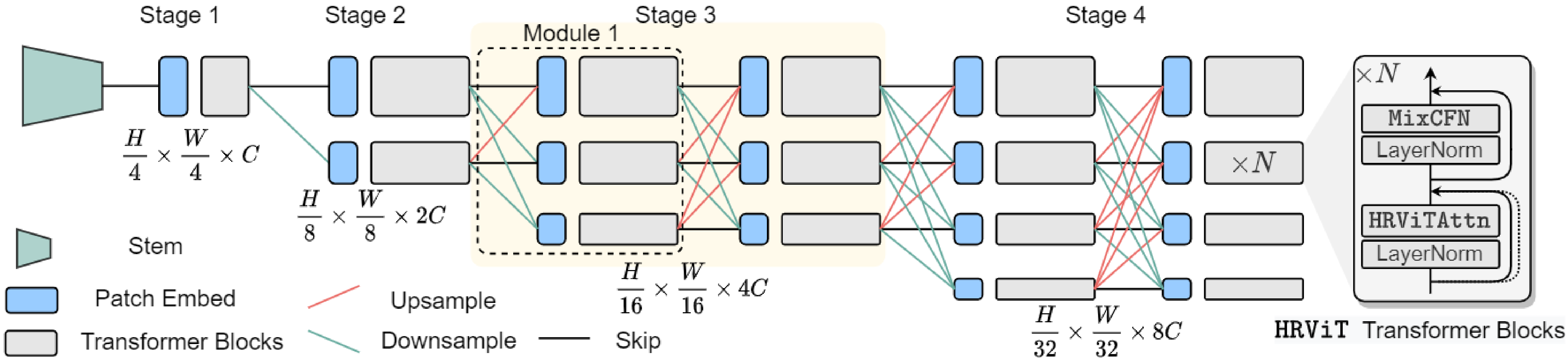
inherit Multi-scale, Cross-resolution .... HR/multi-branch

Large receptive field .... Self-attention

evolve ***High efficiency*** .... **?**

Our Proposed HRViT

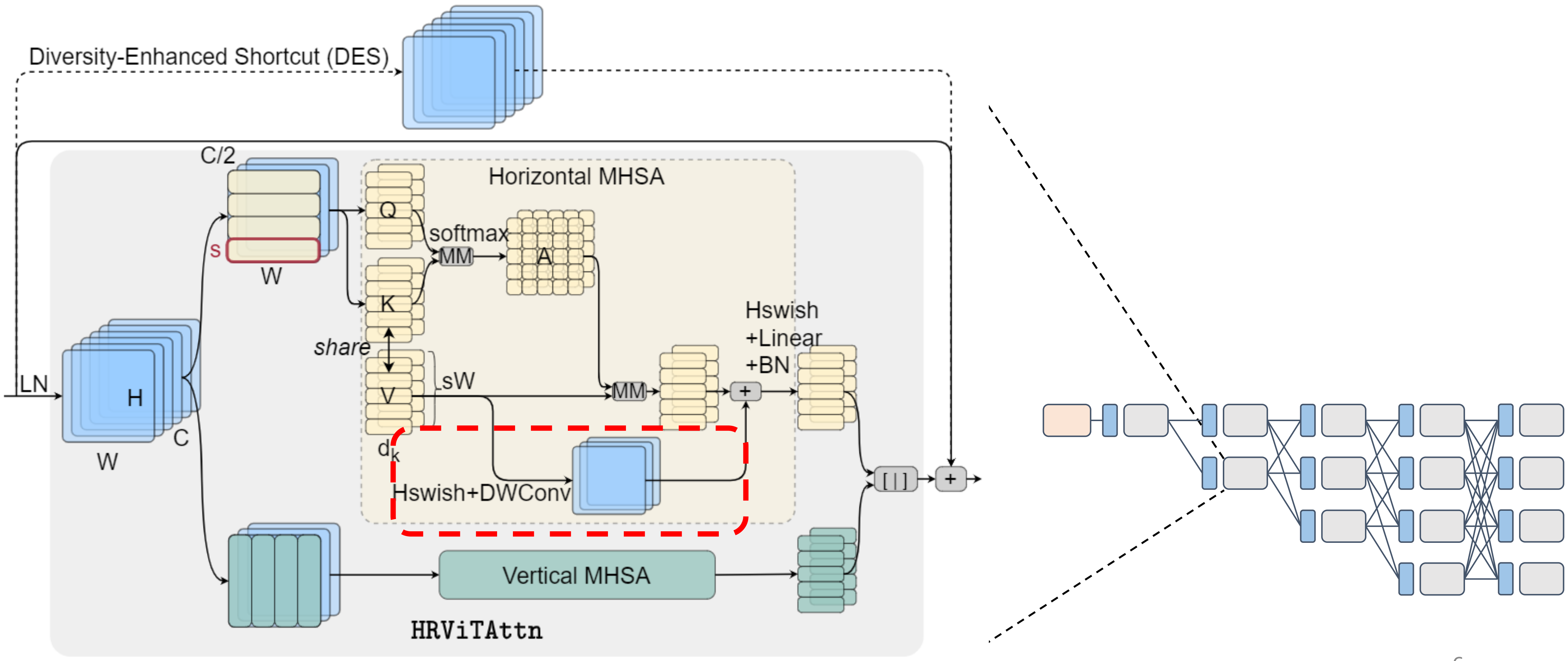
- ***Multi-scale high-resolution vision transformer backbone***
- ***Efficient block-branch co-optimization***
  - **Augmented attentions + Mixed-scale FFN + Cross-resolution fusion + heterogenous branch**
- ***Improved performance-efficiency trade-off***
  - Outperform SoTA SegFormer/CSWin on semantic segmentation



Stage 1
Stage 2
Stage 3
Stage 4
Module 1
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
$\times N$
Stem
Patch Embed
Transformer Blocks
Upsample
Downsample
Skip
$\times N$
MixCFN
LayerNorm
HRViTAttn
LayerNorm
HRViT Transformer Blocks

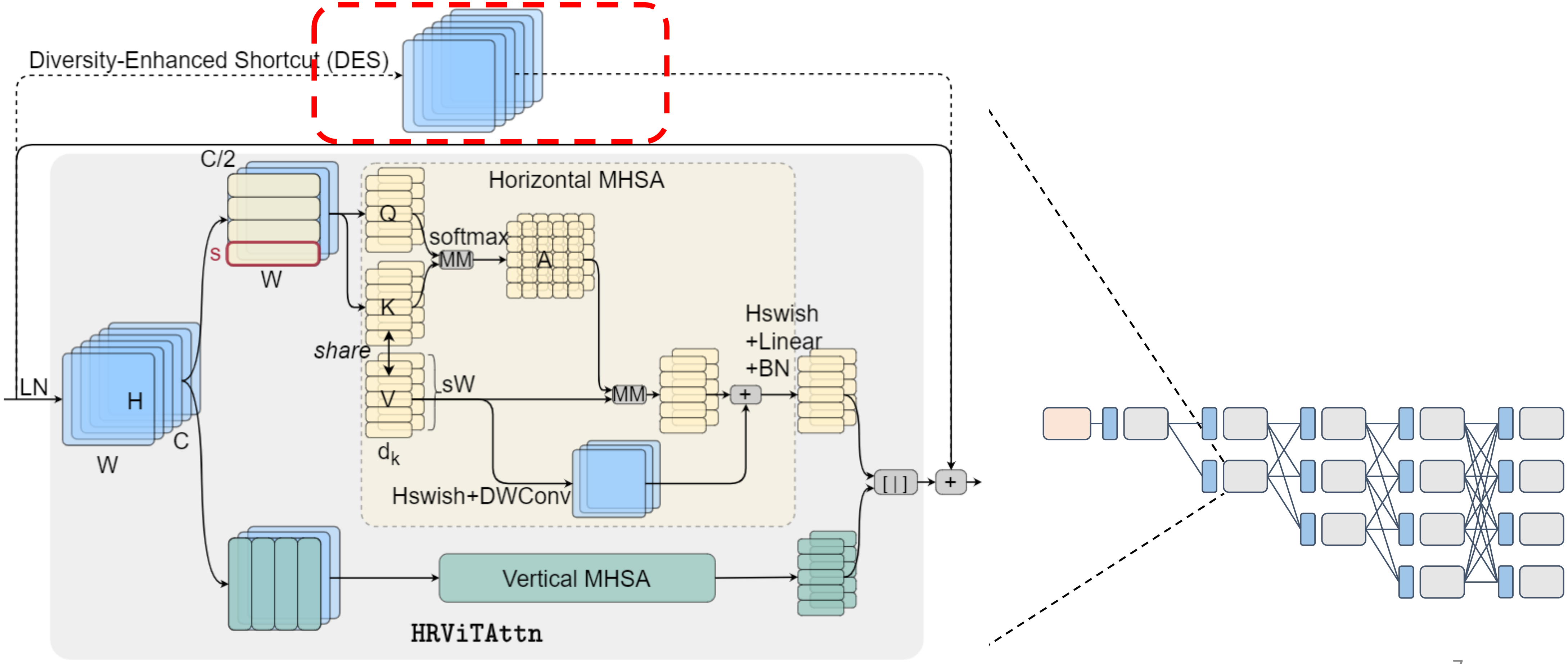
HRViTAttn: Augmented Cross-Shaped Self-Attention

- **Parallel conv** + Diversity-enhanced shortcut
- Share value/key + Augmented nonlinearity



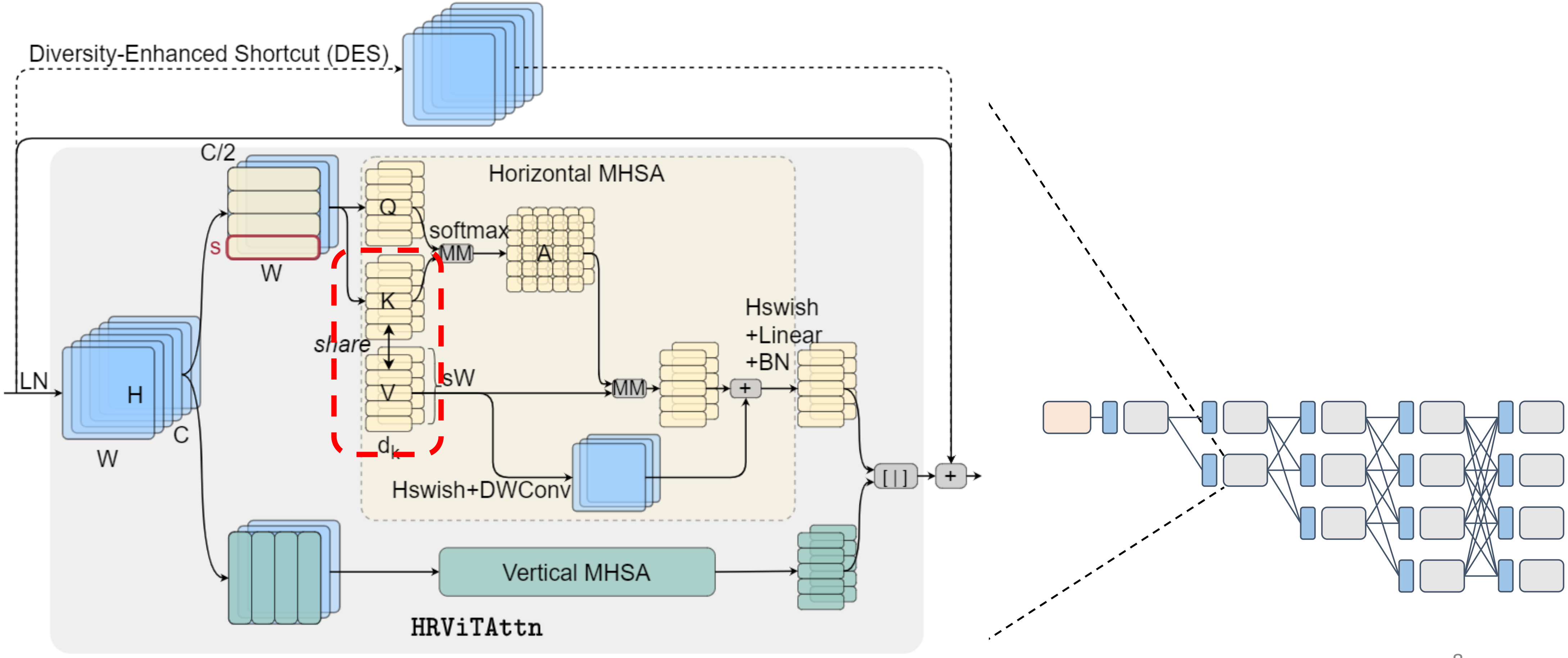
HRViTAttn: Augmented Cross-Shaped Self-Attention

- Parallel conv + **Diversity-enhanced shortcut**
- Share value/key + Augmented nonlinearity



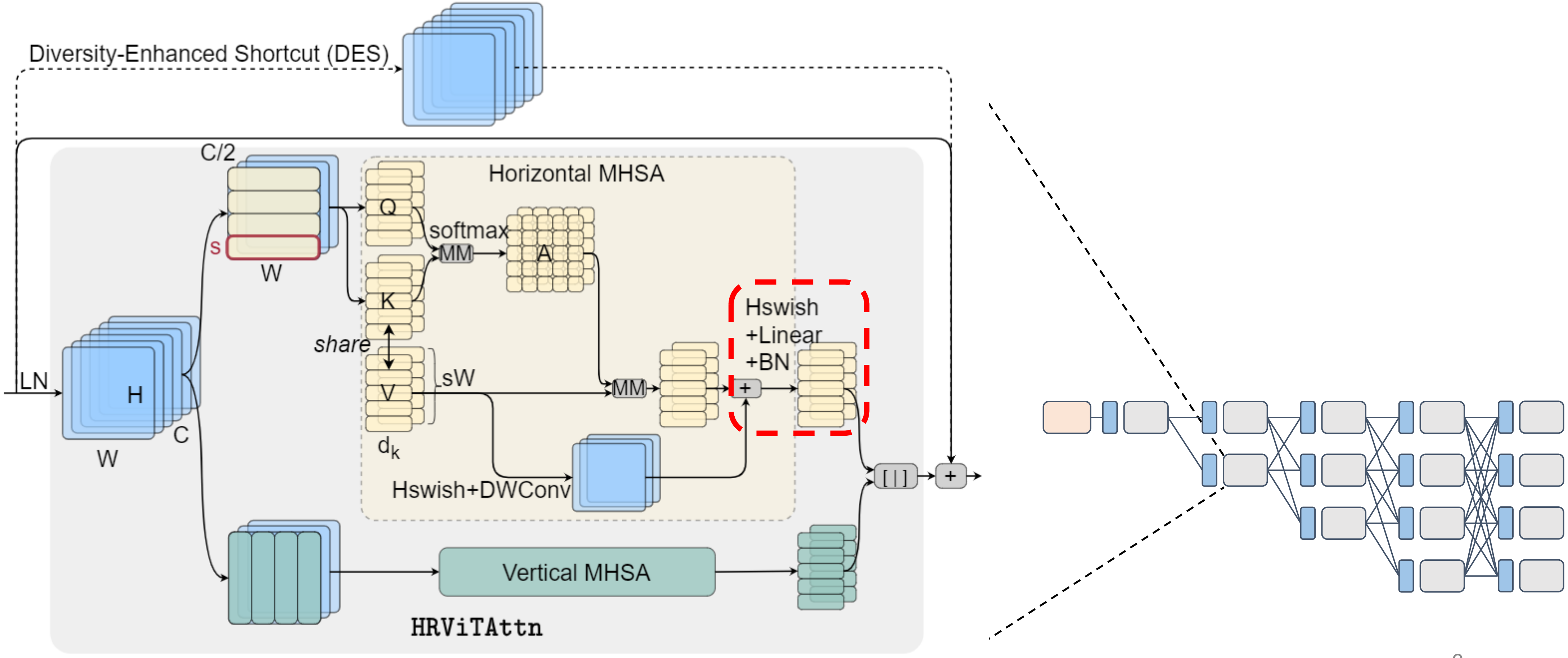
HRViTAttn: Augmented Cross-Shaped Self-Attention

- Parallel conv + Diversity-enhanced shortcut
- **Share value/key** + Augmented nonlinearity



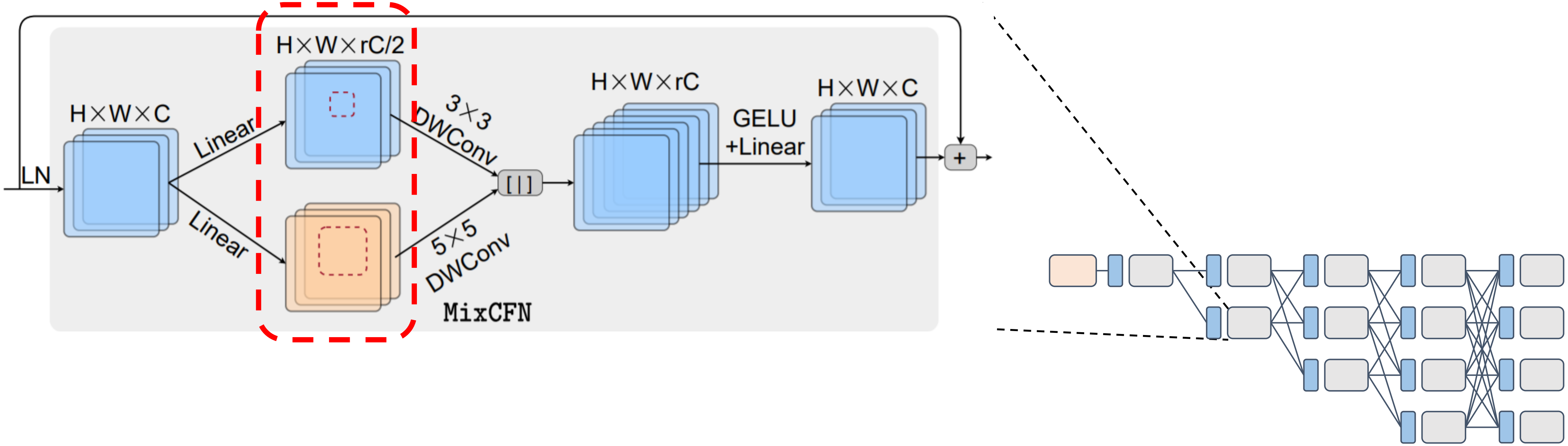
HRViTAttn: Augmented Cross-Shaped Self-Attention

- Parallel conv + Diversity-enhanced shortcut
- Share value/key + **Augmented nonlinearity**



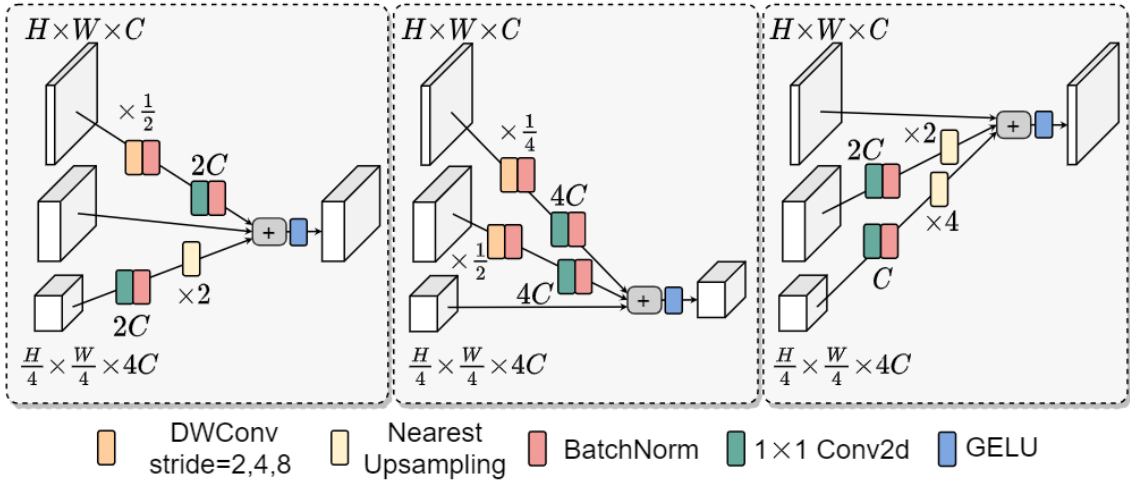
HRViTAttn: Augmented Cross-Shaped Self-Attention

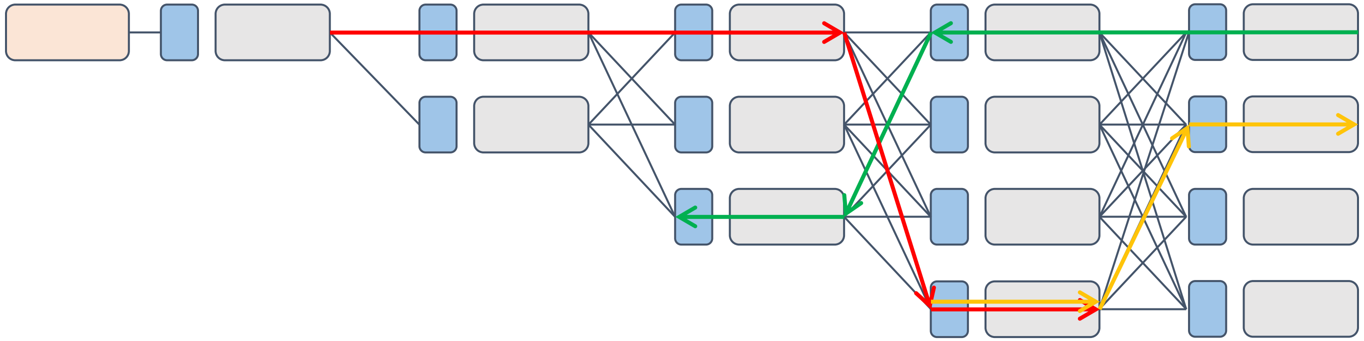
- Mixed-scale depth-wise CONV in FFN
- Reduced expansion ratio



Cross-Resolution Fusion Layer

- ***LR -> HR:*** High-quality HR representation
- ***HR -> LR***: Compensate detailed info loss
- Fortify ***gradient flow*** in deep LR paths
- Lightweight separable CONV





Feed more HR inputs to LR branches

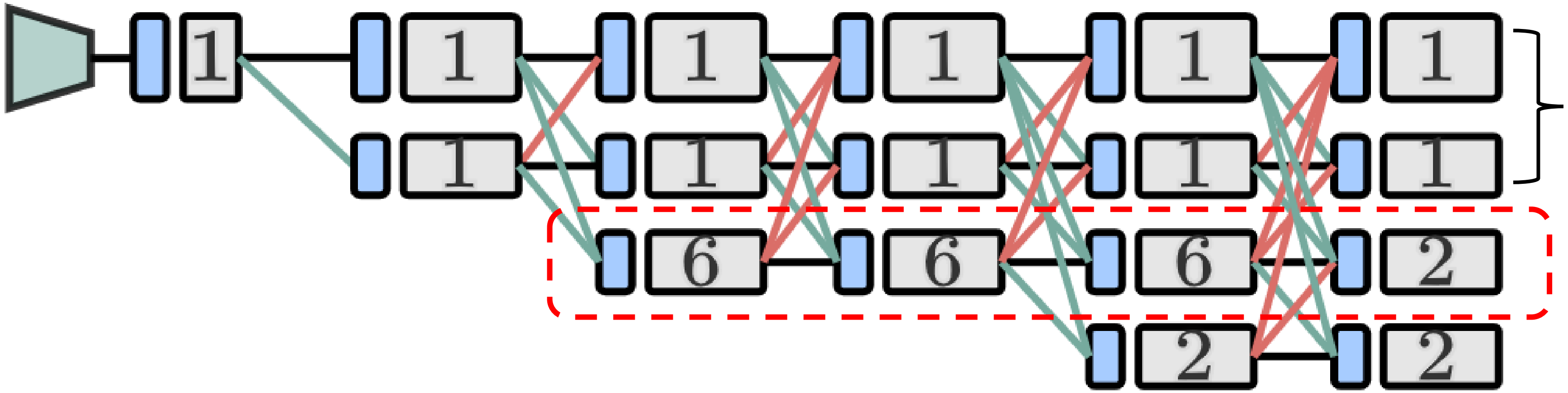
More gradients flow back to deep LR branch

Enlarge receptive fields

Heterogenous Branch Design

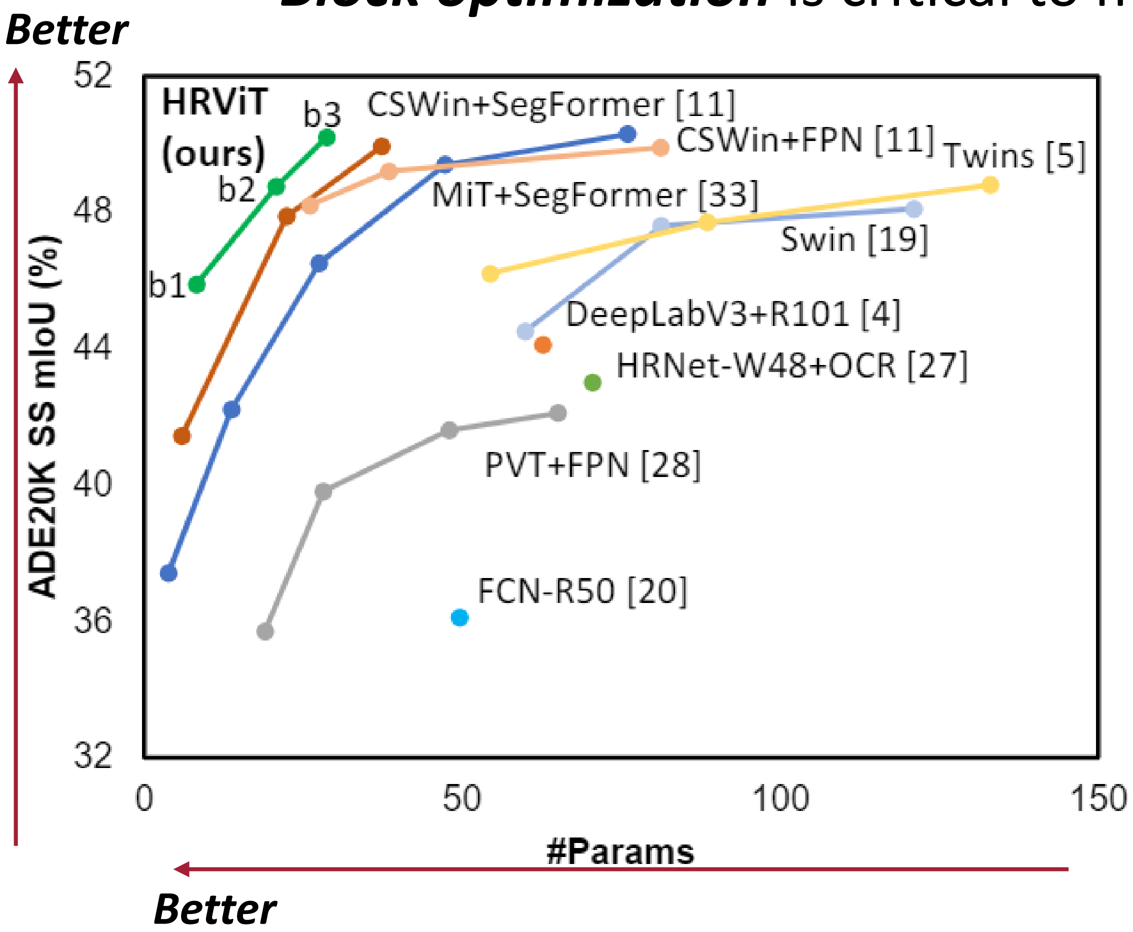
- ***Balance performance & efficiency → key to 'Evolution'***

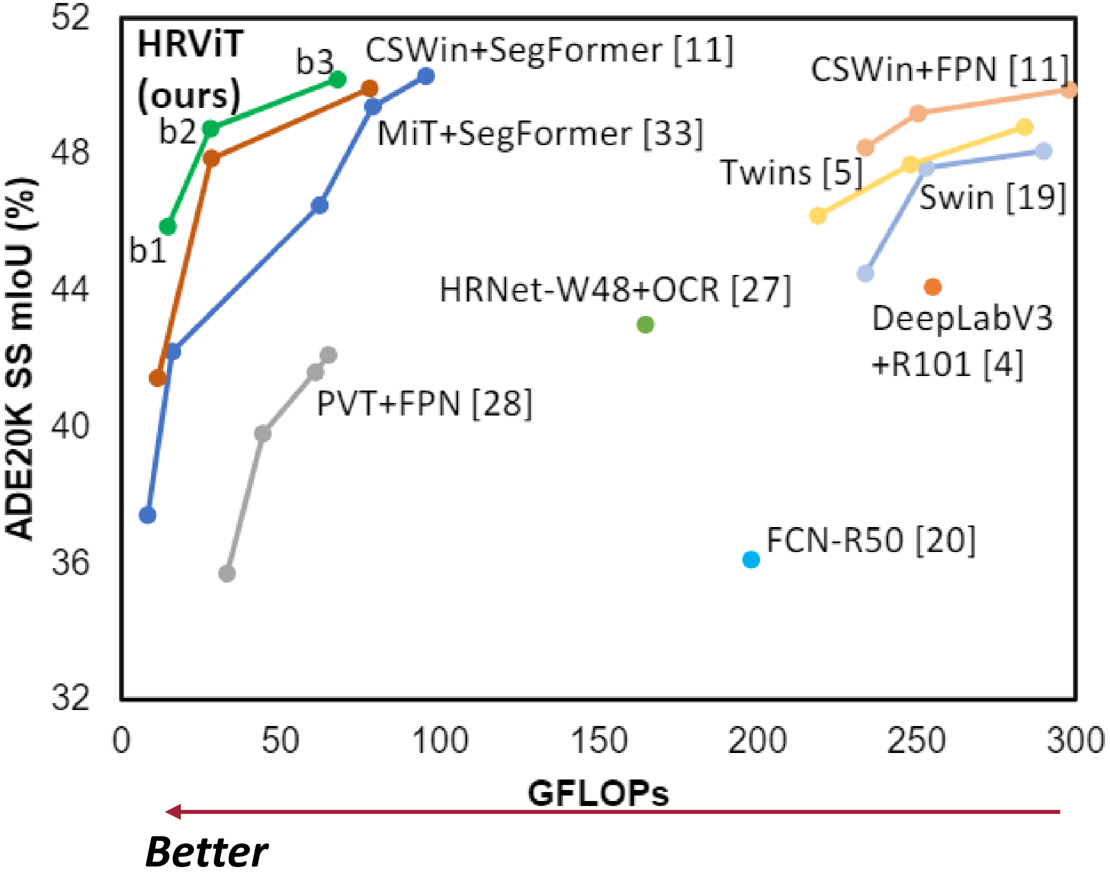
Res.	#Params	FLOPs	Features
HR	Low	Heavy	• **Fine-grained** • **Local**
MR	Mid	Mid	• **High Expressivity**
LR	High	Light	• **Global view**



Main Results on Semantic Segmentation (ADE20k)

- ADE20K/SegFormer Head: **+3.68, +2.26, +0.8** higher mIoU than [MiT, NeurIPS’21]
  - ***HR*** arch brings large ***performance*** gains in *small models*
  - ***Block optimization*** is critical to maintain ***efficiency***



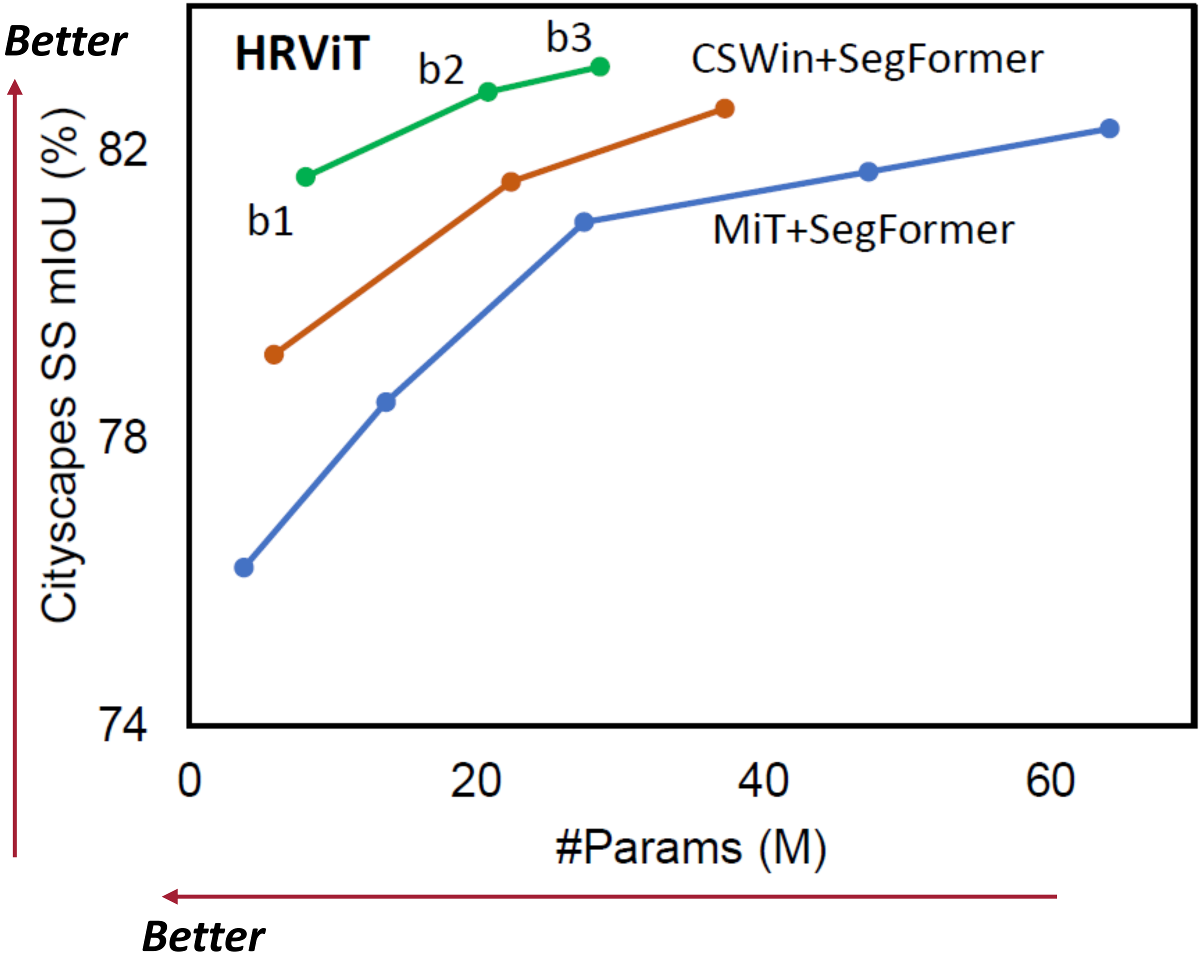


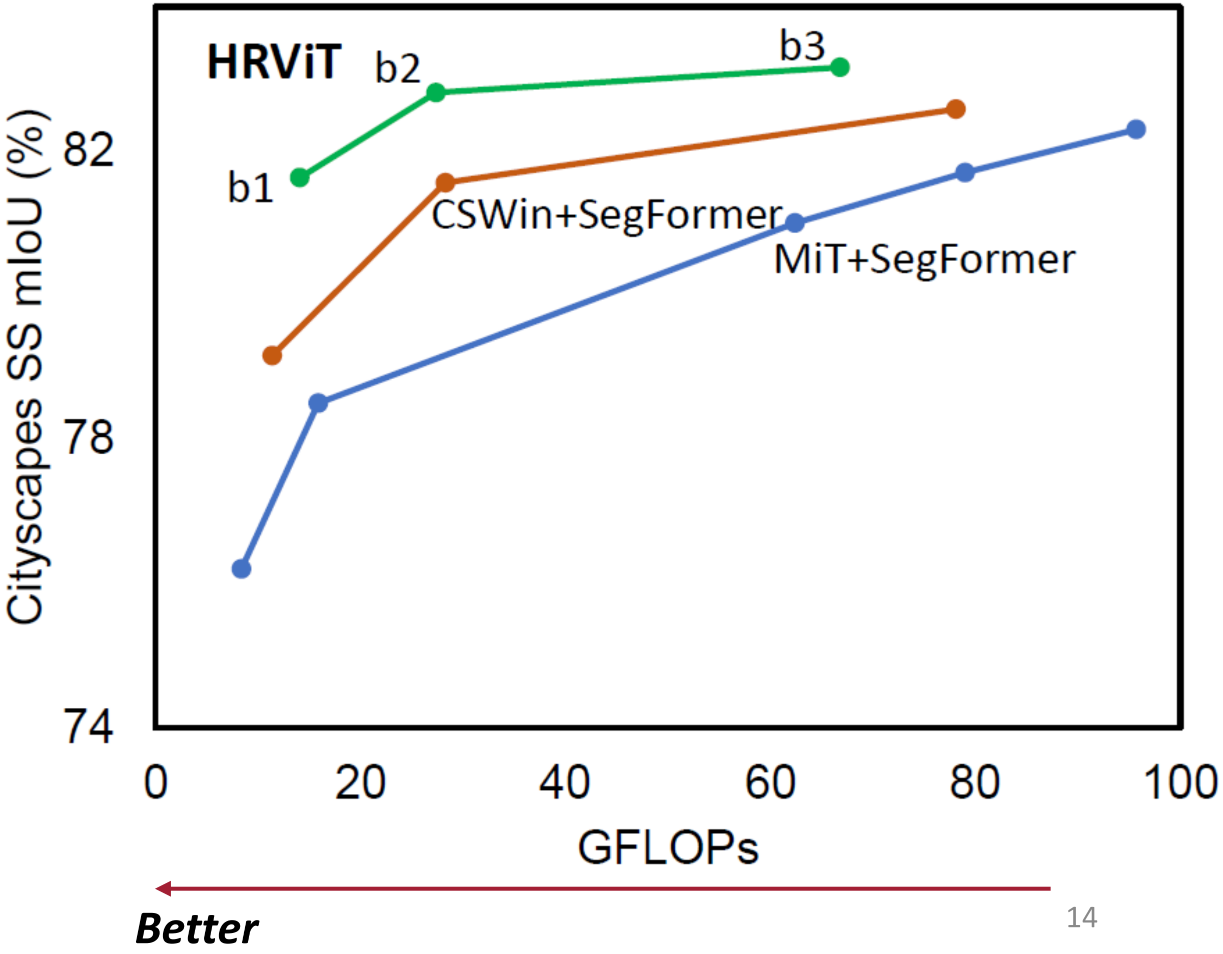
	mIoU
MiT-B1	42.20
CSWin-Ti	41.43
HRViT-b1	**45.88**
MiT-B2	46.50
CSWin-T	47.88
HRViT-b2	**48.76**
MiT-B3	49.40
CSWin-S	49.93
HRViT-b3	**50.20**

SegFormer Head

Main Results on Semantic Segmentation (Cityscapes)

- Cityscapes/SegFormer Head: an average **+2.16 higher mIoU**
- **30.7% fewer params** + **23.1% less computation**





HRViT: Take-aways

- **HR + ViT**
  - HR architecture makes ViTs stronger semantic segmentation backbones

Multi-scale

- **HR > Seq**
  - HR multi-scale architecture outperforms sequential counterparts

Cross-resolution

- **Optimized ViT blocks > Original ViT blocks**
  - Careful block optimization is critical to balanced efficiency and performance

Efficiency Opt.

- **Customized HR Arch > Original HR Arch**
  - Heterogeneous branch design is important to efficiency-accuracy trade-off

Customization

Thank you

Q & A