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]The University of Texas at Austin, [2]Meta Platforms Inc.

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

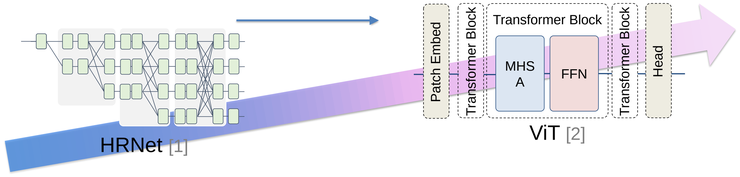
Poster 3.1 143a

Introduction & Motivation

➢ **Performance** vs. **Efficiency** trade-off in **dense vision tasks**

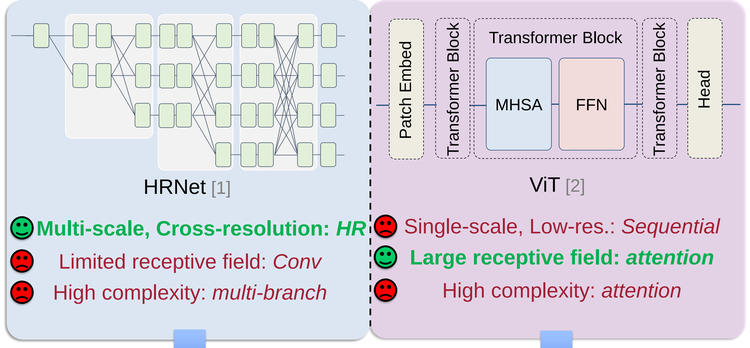
- **Low hardware cost on edge devices**
- **High-performance detection, segmentation, etc.**

➢ **CNN** **Vision Transformer**



➢ How to ***inherit*** from **HRNet** and **ViTs** and ***evolve***?

Directly merge HRNet + ViT: **prohibitive cost**



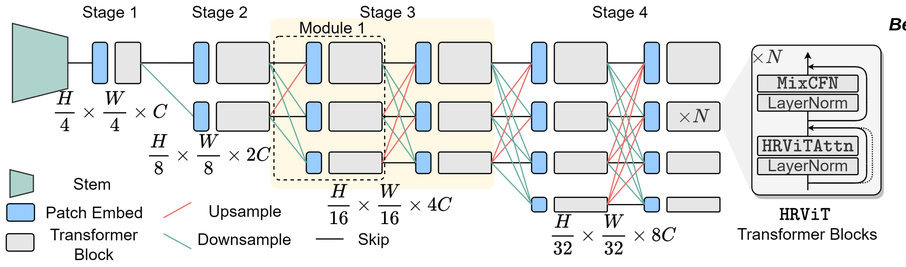
inherit Multi-scale, Cross-resolution ..... HR/multi-branch

Large receptive field ..... Self-attention

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

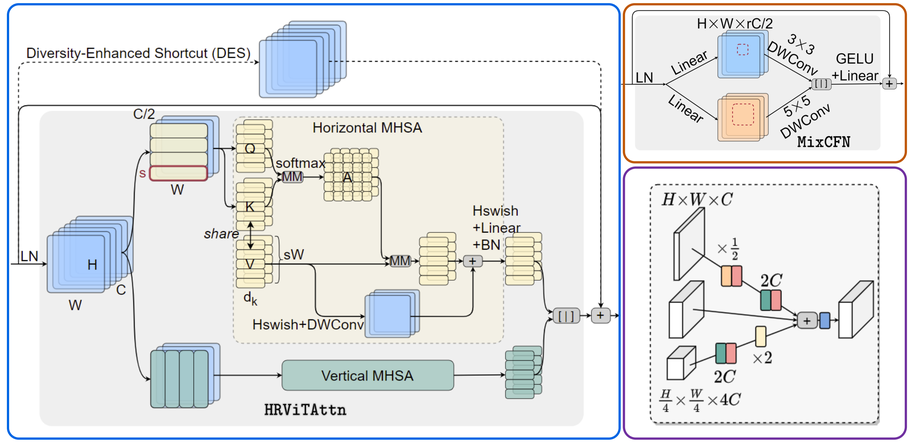
Need **synergistic customization**

Proposed HRViT Architecture



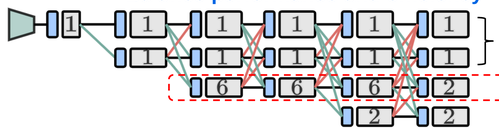
➢ ***Efficient block optimization***

- **Augmented attentions** + **Mixed-scale FFN** + **Cross-resolution fusion**



➢ ***Heterogenous branch design***

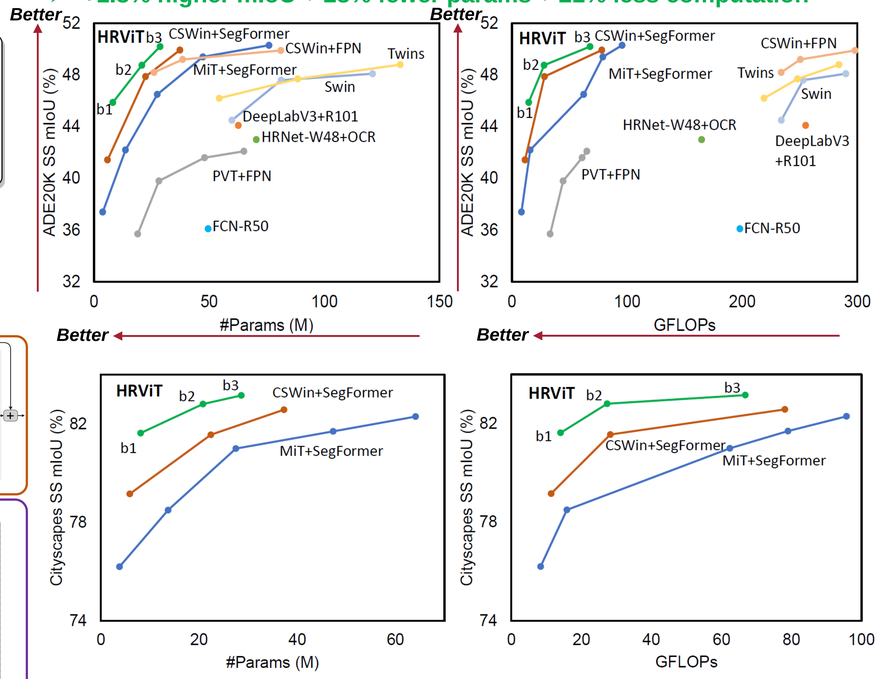
- Customized window size, branch depth, MLP ratio, #channels, etc
- **Balance performance and efficiency**



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

Semantic Segmentation on ADE20K / Cityscapes

➢ **+1.8% higher mIoU + 28% fewer params + 21% less computation**



Conclusion & Future Direction

➢ ***Multi-scale representation learning*** is critical to ViTs

➢ ***Co-optimization*** is the key to balancing performance and efficiency

➢ Extend to more dense prediction vision tasks + deploy to **edge** devices

Open-source: github.com/facebookresearch/HRViT

References

[1] J. Wang, et al., "Deep high-resolution representation learning for visual recognition," TPAMI, 2019.
[2] A. Dosovitskiy, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
[3] E. Xie, et al., "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," NeurIPS, 2021.
[4] Z. Liu, et al., Swin Transformer: Hierarchical Vision Transformer using Shifted Windows"," ICCV, 2021.