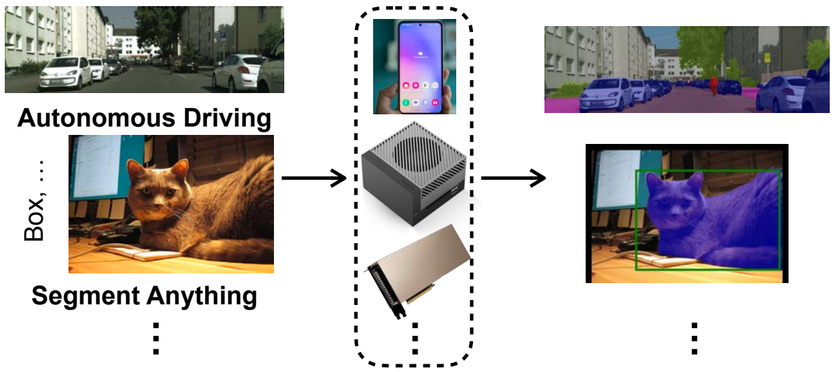
EfficientViT: Multi-Scale Linear Attention for High-Resolution Dense Prediction

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, Song Han
MIT, Zhejiang University, Tsinghua University, MIT-IBM Watson AI Lab

ICCV23 PARIS

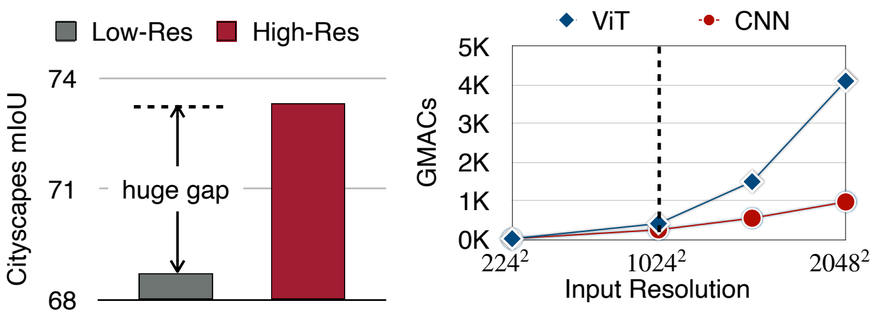
Efficient High-Resolution Dense Prediction on Hardware



Enable efficient high-resolution dense prediction on hardware can benefit many use cases.

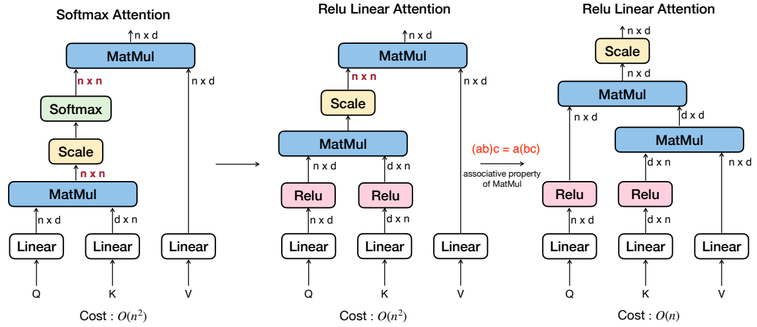
Challenge: Apply Transformer to High-Resolution Images

Details and small objects are missing in low-resolution images.

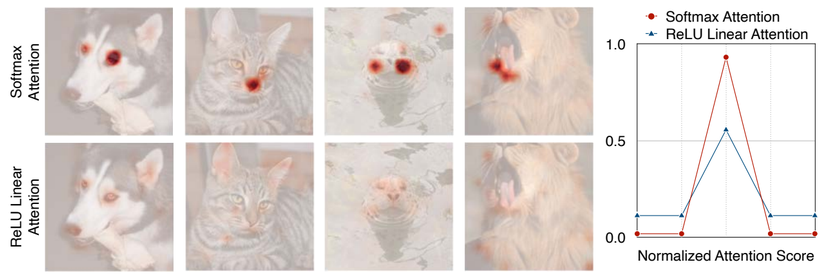


- High resolution is essential for achieving good performances in dense prediction tasks (e.g., segmentation).
- ViT's computational cost grows quadratically as the input resolution increases.

EfficientViT: Lightweight ViT architecture for high-resolution dense prediction

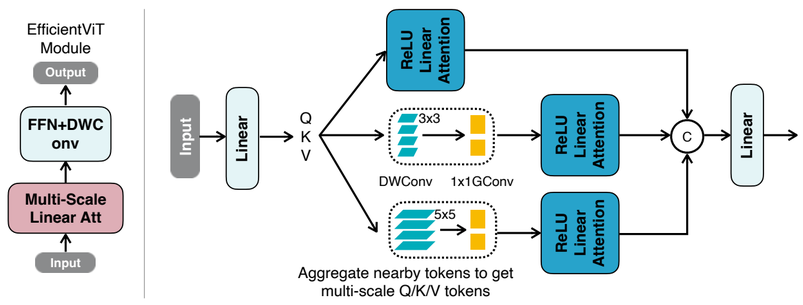


Enable global receptive field with lightweight ReLU linear attention.



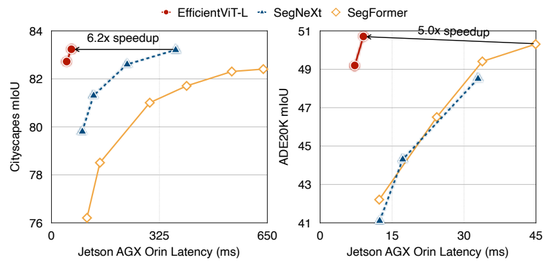
- But there is no free lunch. ReLU linear attention cannot produce sharp distributions. Thus, it is good at capturing global context information but bad at capturing local information.
- It also lacks multi-scale learning ability.

EfficientViT: Multi-Scale Linear Attention



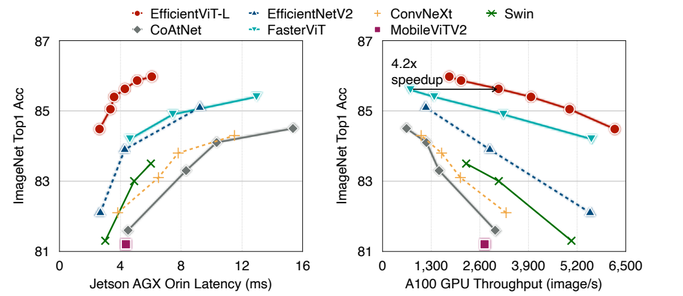
- Linear computational complexity.
- Global receptive field.
- Multi-scale learning.
- No hardware-unfriendly operations.

Results on Segmentation, Classification, Segment Anything



Demo Video

Significant speedup on Jetson AGX Orin without performance drop.



Efficiency boost on both edge and cloud GPUs.

84x speedup on A100 GPU while maintaining comparable zero-shot image segmentation quality.

MIT HAN LAB
Hardware, AI and Neural-nets