EfficientViT: Lightweight Multi-Scale Attention for On-Device Semantic Segmentation

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, Song Han
Massachusetts Institute of Technology

Efficient Semantic Segmentation on Edge Device

Challenges:

- **large gap** between the computational cost required by SOTA semantic segmentation models and the limited resources of edge devices.
- semantic segmentation is a dense prediction task requiring **high-resolution images** and **strong context information extraction ability** to deliver good performances

Features	SegFormer [45]	HRFormer [49]	SegNeXt [17]	**EfficientViT**
Global receptive field	✓			✓
Multi-scale learning		✓	✓	✓
Linear computational complexity		✓	✓	✓
Hardware efficiency				✓

- Limitation of prior semantic segmentation models

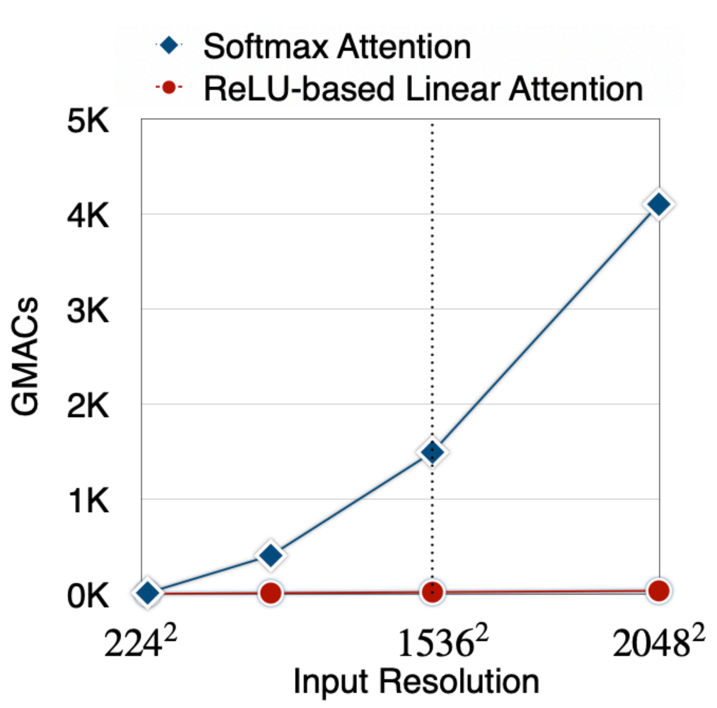
Lightweight Multi-Scale Attention: Trade Slight Capacity Loss for Significant Efficiency Boost

- Replace the similarity function in attention

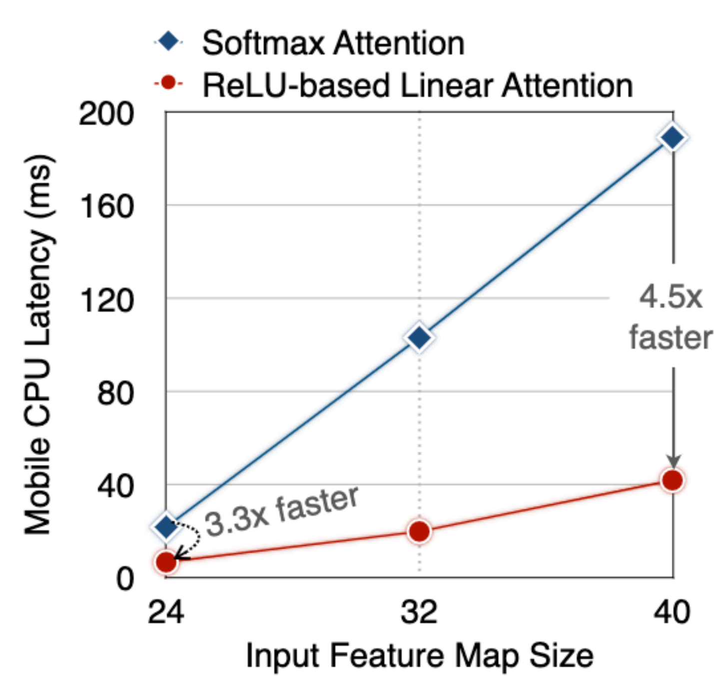
$$Sim(Q,K) = \exp\left(\frac{QK^T}{\sqrt{d}}\right) \longrightarrow Sim(Q,K) = \phi(Q)\phi(K)^T = ReLU(Q)ReLU(K)^T$$

- Change the order of matrix multiplication without changing functionality

$$O_i = \frac{\sum_{j=1}^{N}[\mathrm{ReLU}(Q_i)\mathrm{ReLU}(K_j)^T]V_j}{\mathrm{ReLU}(Q_i)\sum_{j=1}^{N}\mathrm{ReLU}(K_j)^T} = \frac{\sum_{j=1}^{N}\mathrm{ReLU}(Q_i)[(\mathrm{ReLU}(K_j)^TV_j)]}{\mathrm{ReLU}(Q_i)\sum_{j=1}^{N}\mathrm{ReLU}(K_j)^T} = \frac{\mathrm{ReLU}(Q_i)(\sum_{j=1}^{N}\mathrm{ReLU}(K_j)^TV_j)}{\mathrm{ReLU}(Q_i)(\sum_{j=1}^{N}\mathrm{ReLU}(K_j)^T)}$$

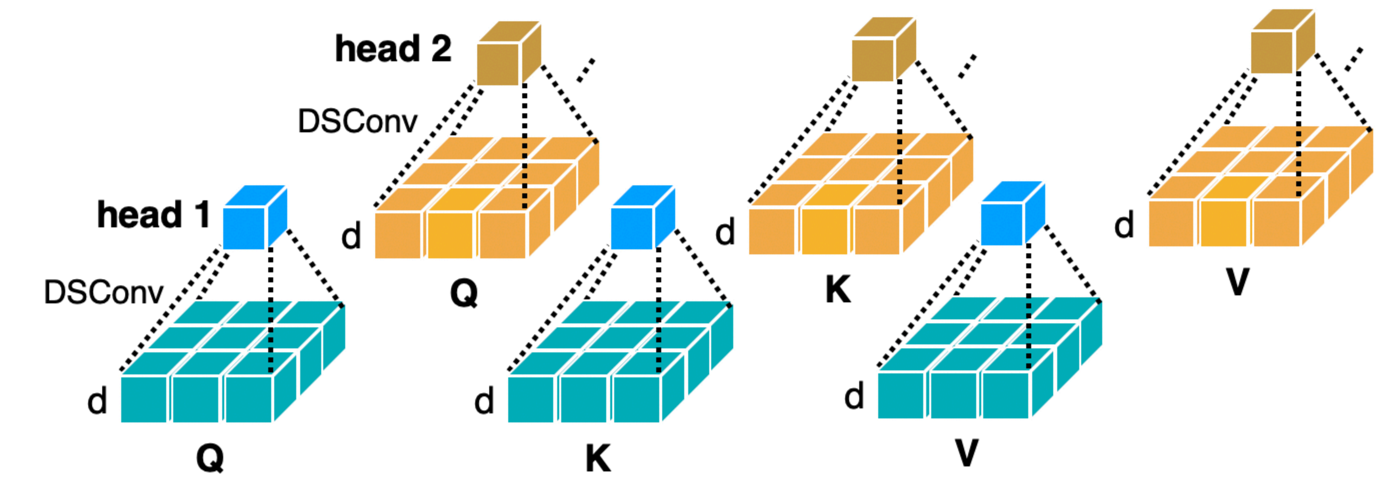


- Lower computational complexity



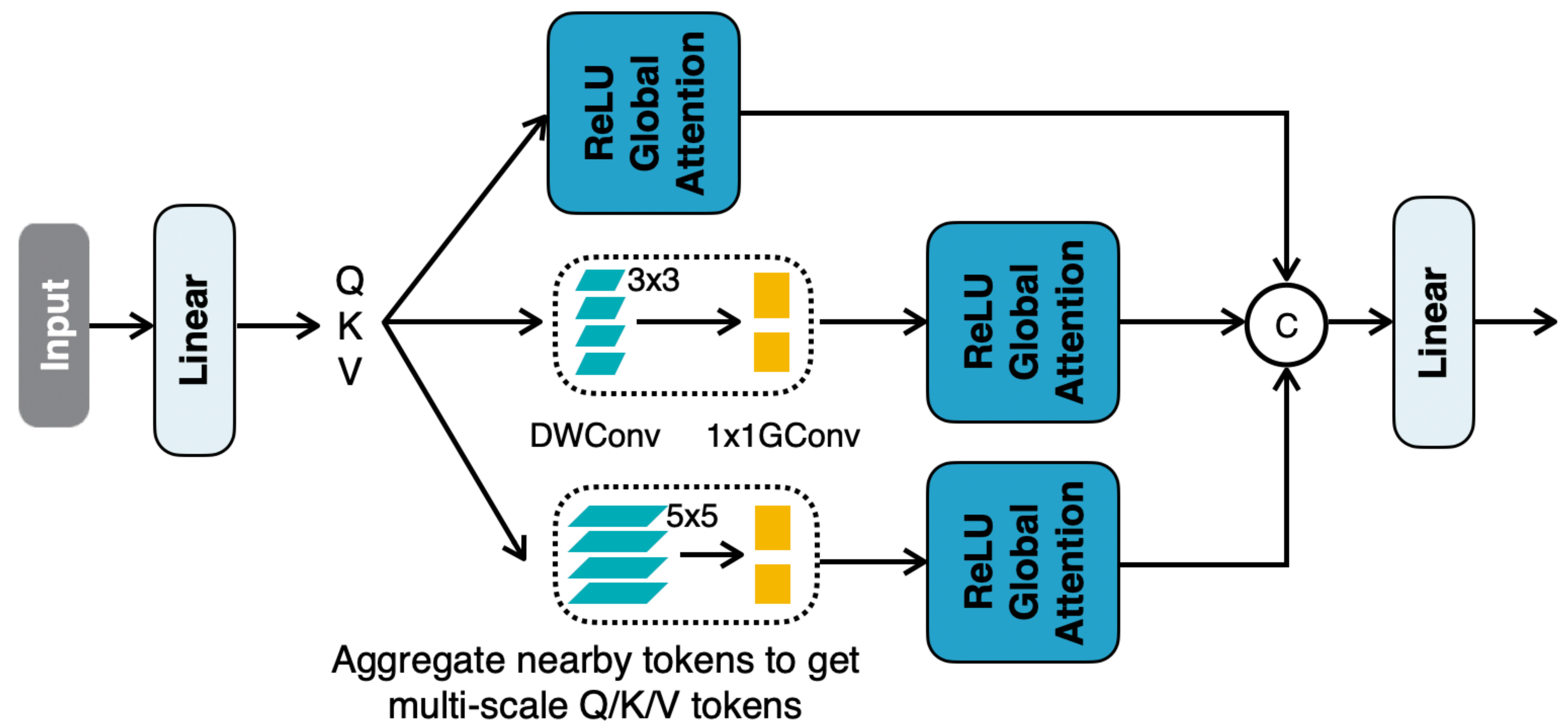
- Better hardware efficiency

Lightweight Multi-Scale Attention: Generate Multi-Scale Tokens



- Aggregate nearby tokens to generate multi-scale tokens

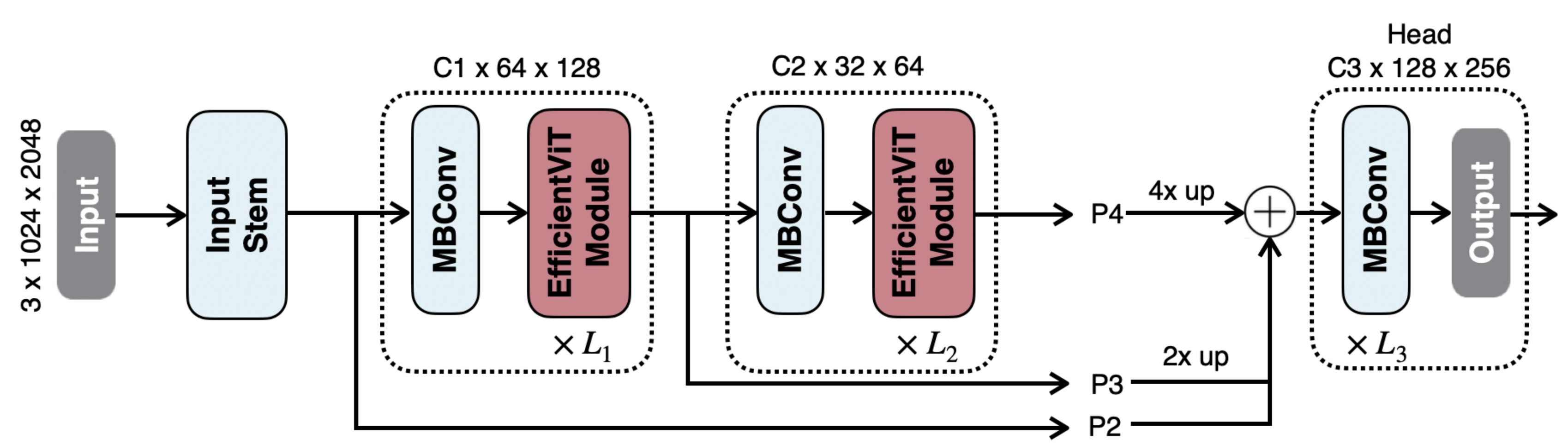
Lightweight Multi-Scale Attention: Block Design



Components		mIoU ↑	Params ↓	MACs ↓
Multi-scale	Global att.			
		68.1	0.7M	4.4G
✓		72.3	0.7M	4.4G
	✓	72.2	0.7M	4.4G
✓	✓	**74.5**	0.7M	4.4G

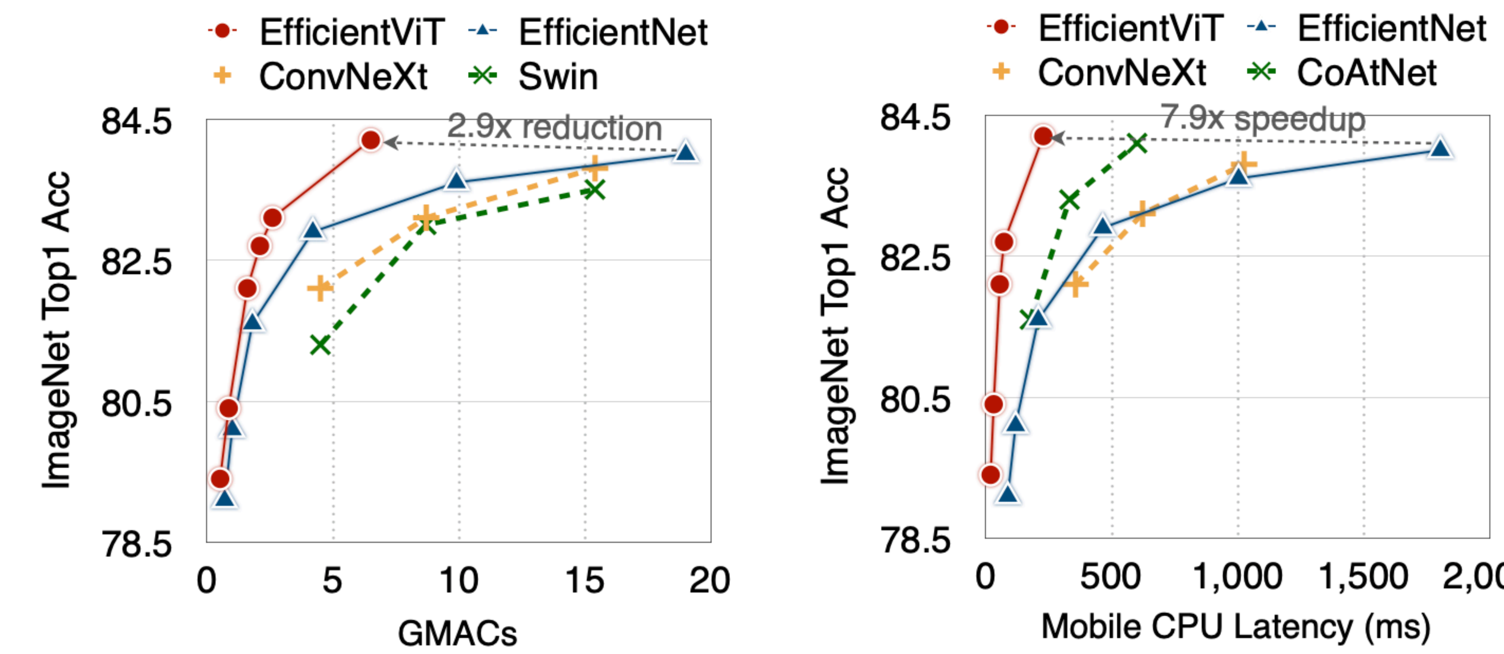
- Both global receptive field and multi-scale learning are essential for obtaining good semantic segmentation performance.

EfficientViT Macro Architecture



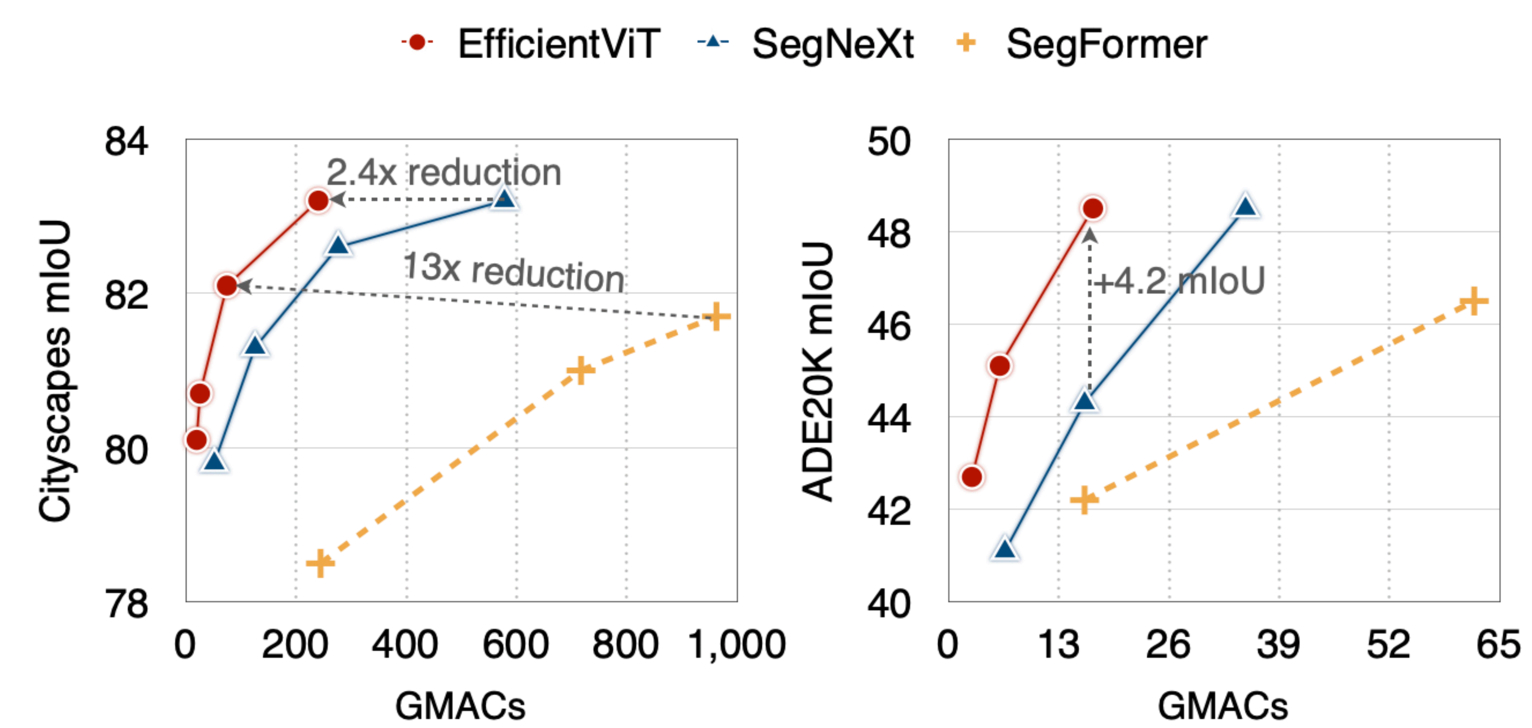
- Backbone: strong context information extraction capacity.
- Head: simple and lightweight.

Backbone Results on ImageNet

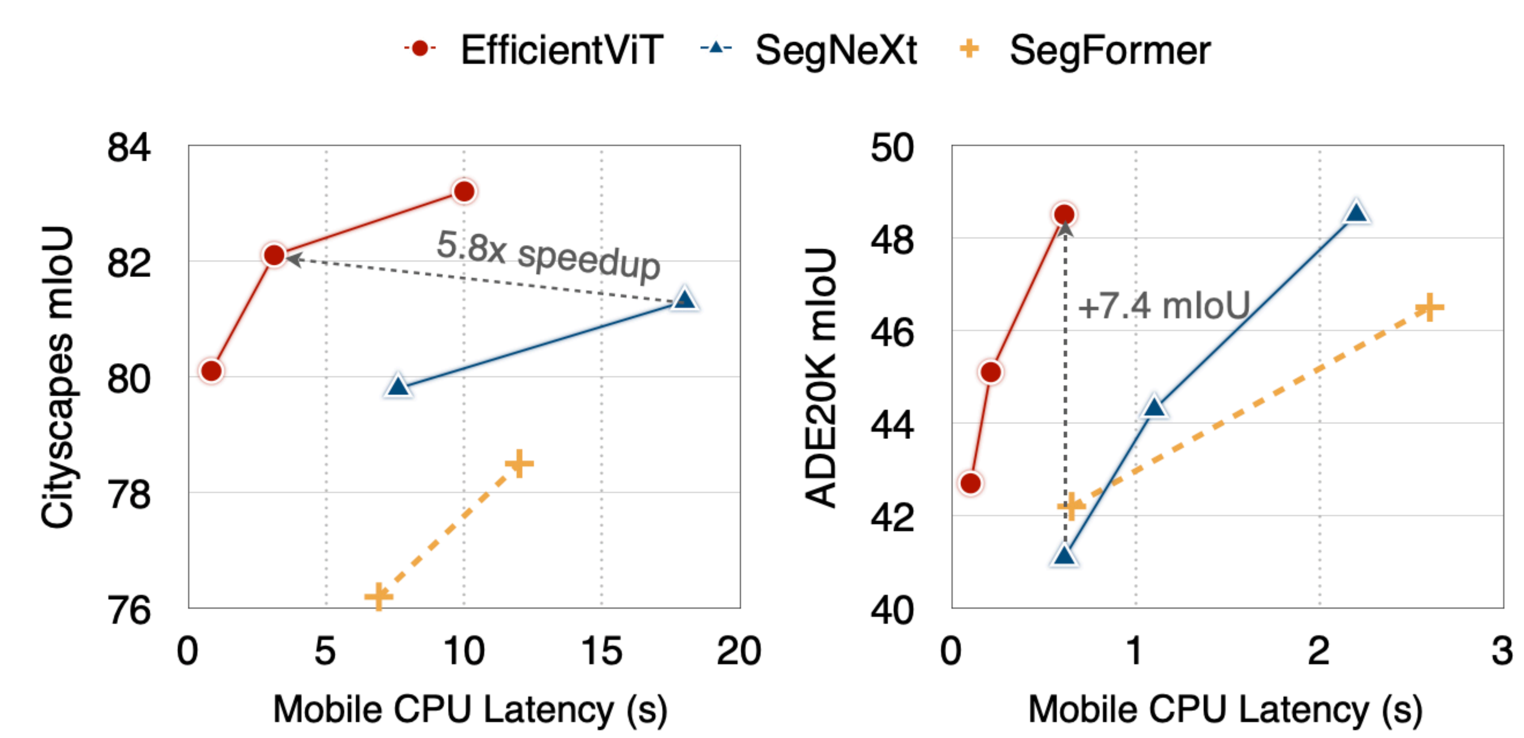


- **2.9x** MACs reduction without performance loss on ImageNet compared with EfficientNet-B6.
- **7.9x measured speedup** on Qualcomm Snapdragon 8Gen1 CPU over EfficientNet-B6 without accuracy loss.

Semantic Segmentation Results



- Cityscapes: **13x** and **2.4x** MACs reduction over SegFormer and SegNeXt.
- ADE20K: **4.2 mIoU gain** over SegNeXt.



- Cityscapes: **5.8x measured speedup and higher mIoU** than SegNeXt.
- ADE20K: **7.4 mIoU gain** over SegNeXt with the same latency.