

Lite Pose: Efficient Architecture Design for 2D Human Pose Estimation

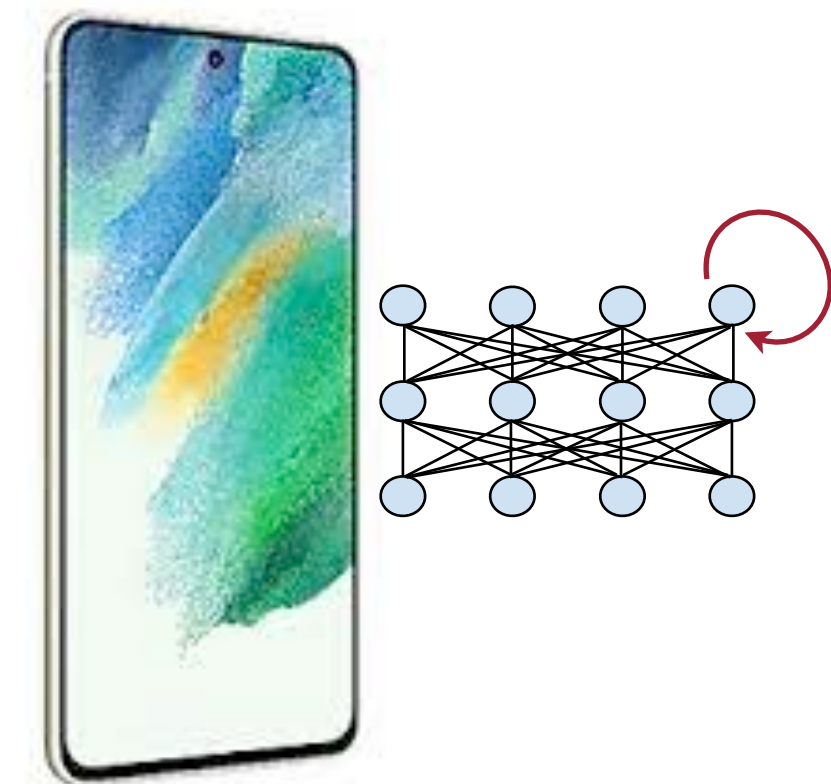
Yihan Wang¹, Muyang Li², Han Cai³, Wei-Ming Chen³, Song Han³

¹Tsinghua University ²CMU ³MIT

Real-Time Multi-Person Pose Estimation on Edge



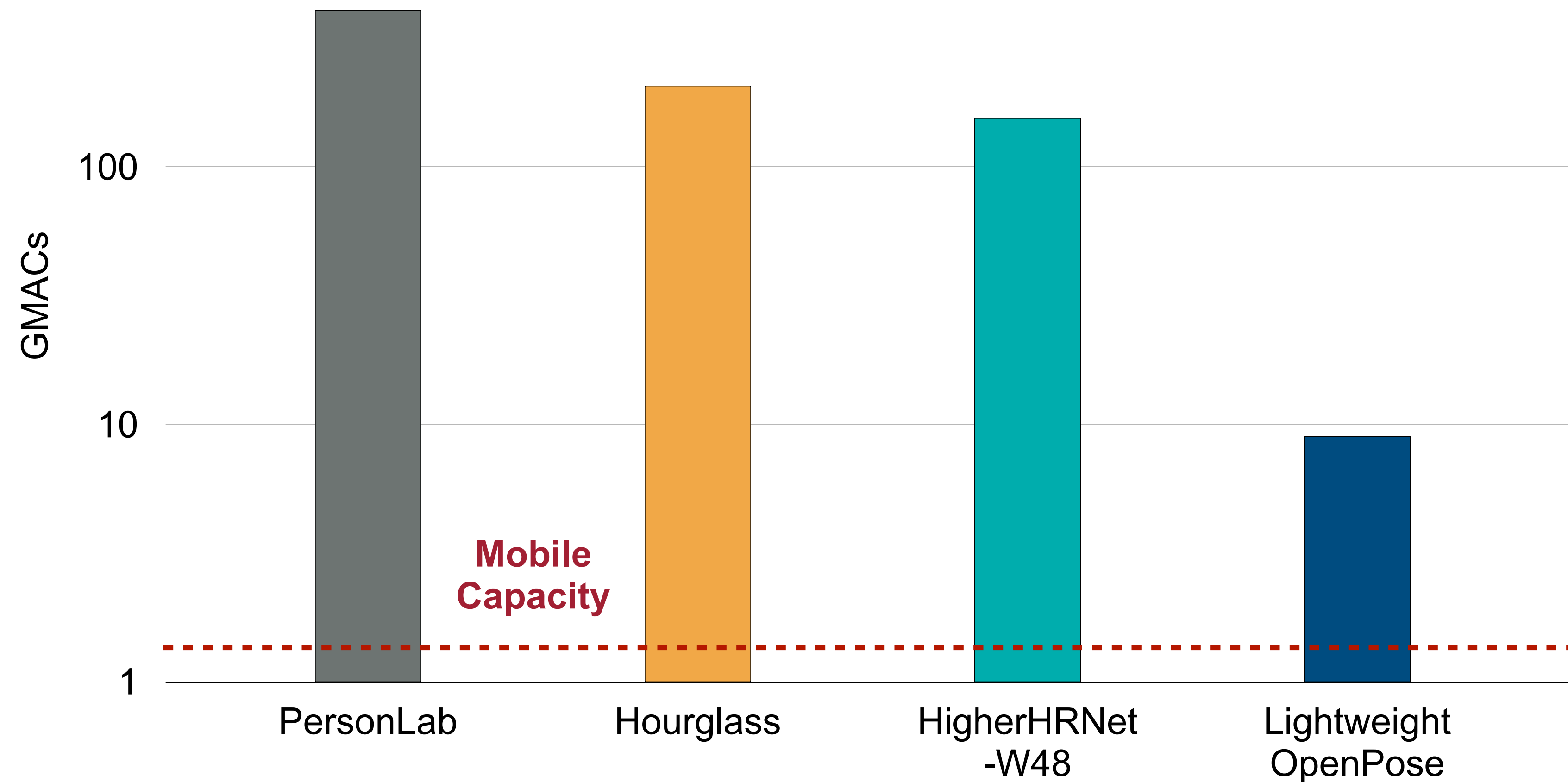
Multi-Person
Pose Estimation



Edge Devices

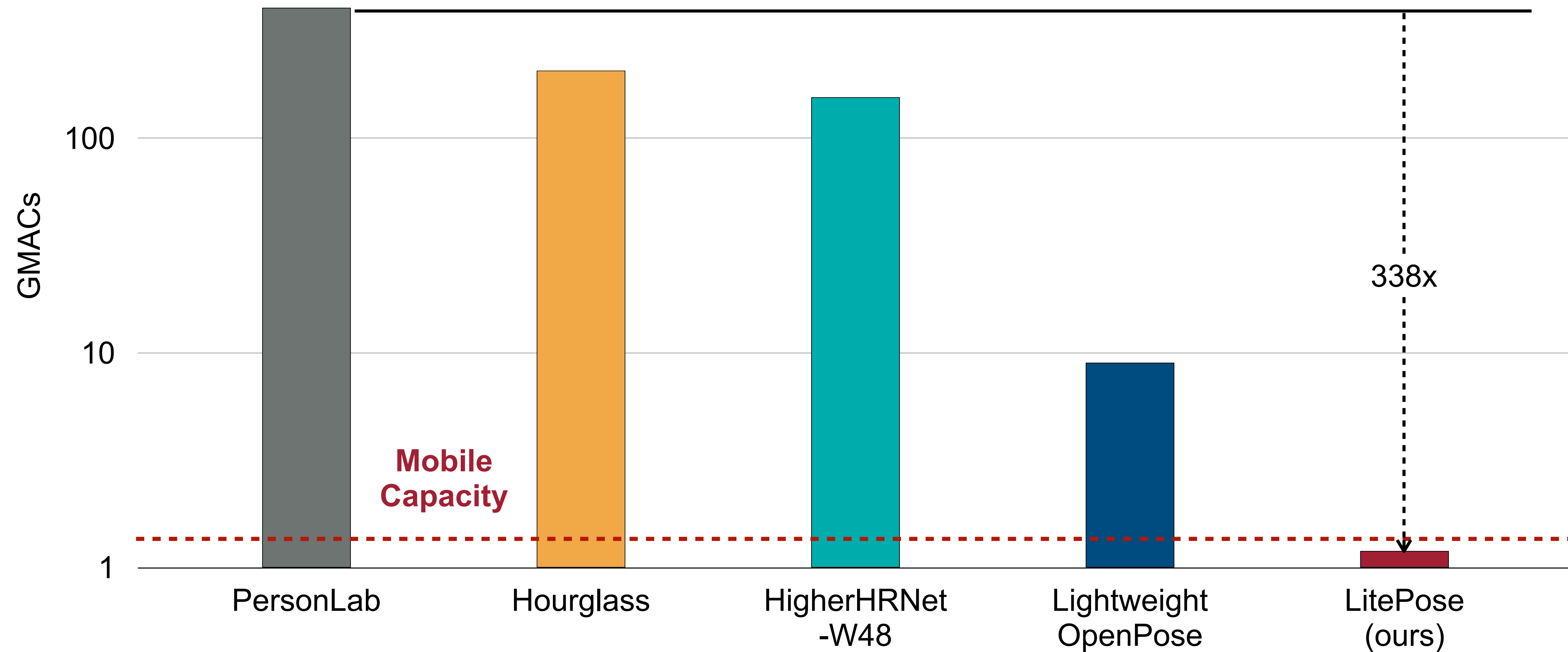
Many human-centered vision applications rely on **real-time multi-person** pose estimation on **edge** devices, requiring **low-computation** pose estimation models.

Current Pose Estimation Models are too Heavy for Edge Devices



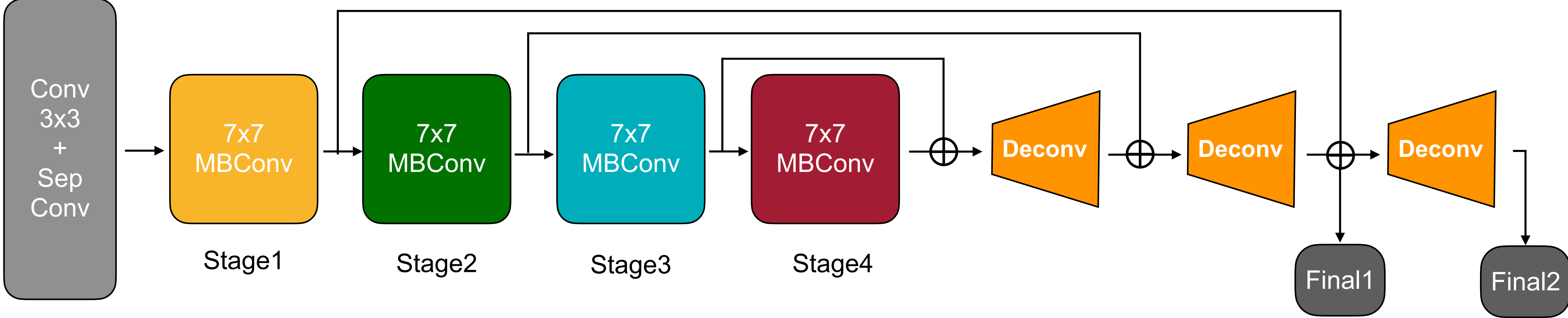
However, current pose estimation models are too **heavy** for edge devices.

Current Pose Estimation Models are too Heavy for Edge Devices



However, current pose estimation models are too **heavy** for edge devices. We introduce **LitePose** to close the gap.

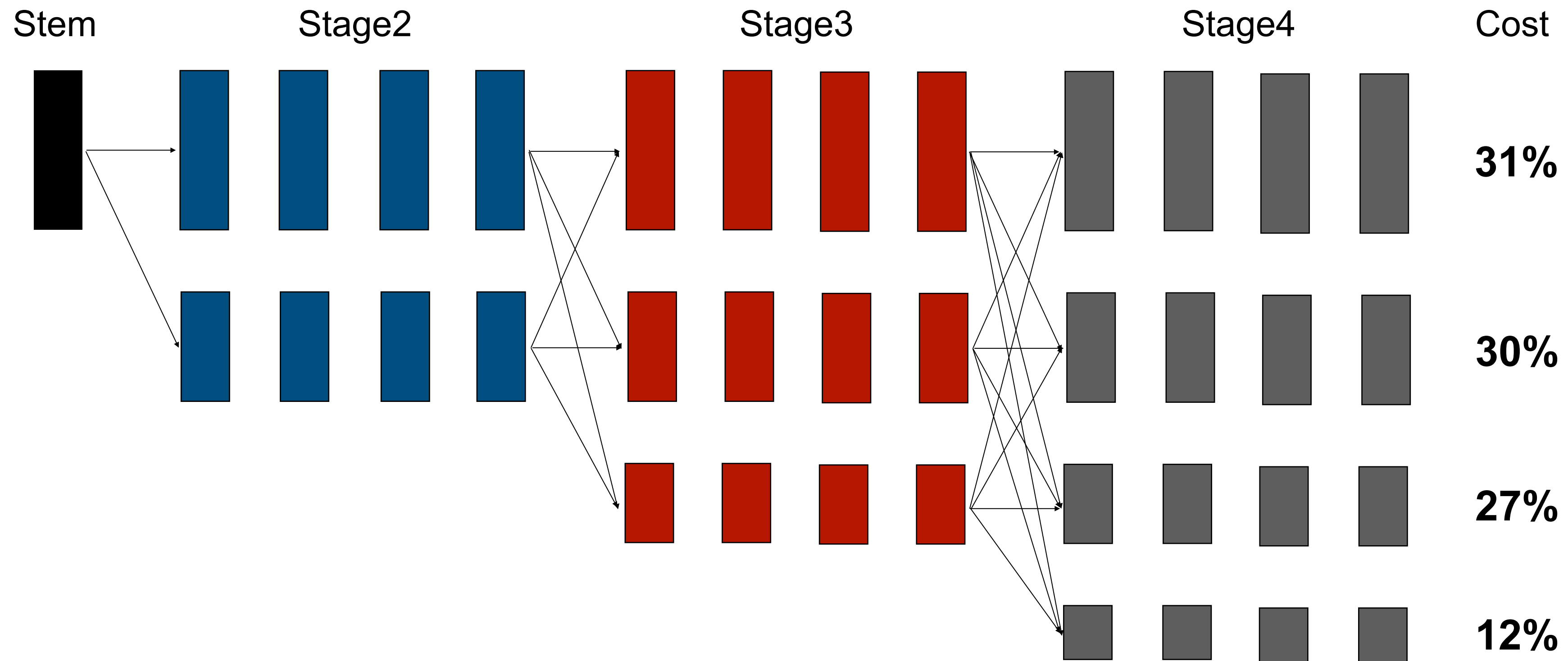
Overview of LitePose



Key insights:

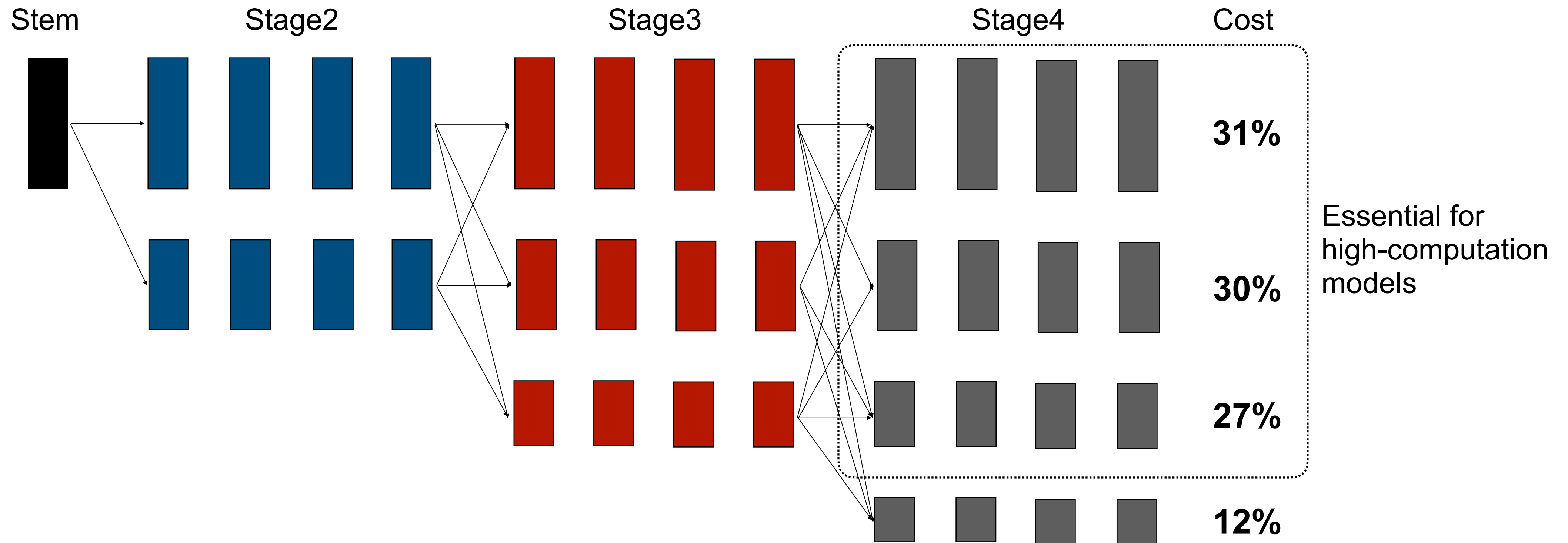
- 1. Single-branch architecture is efficient
- 2. Large kernel convolution is efficient.
- 3. Light-weight fusion deconv head.

High-Resolution Branches are the Key Bottleneck



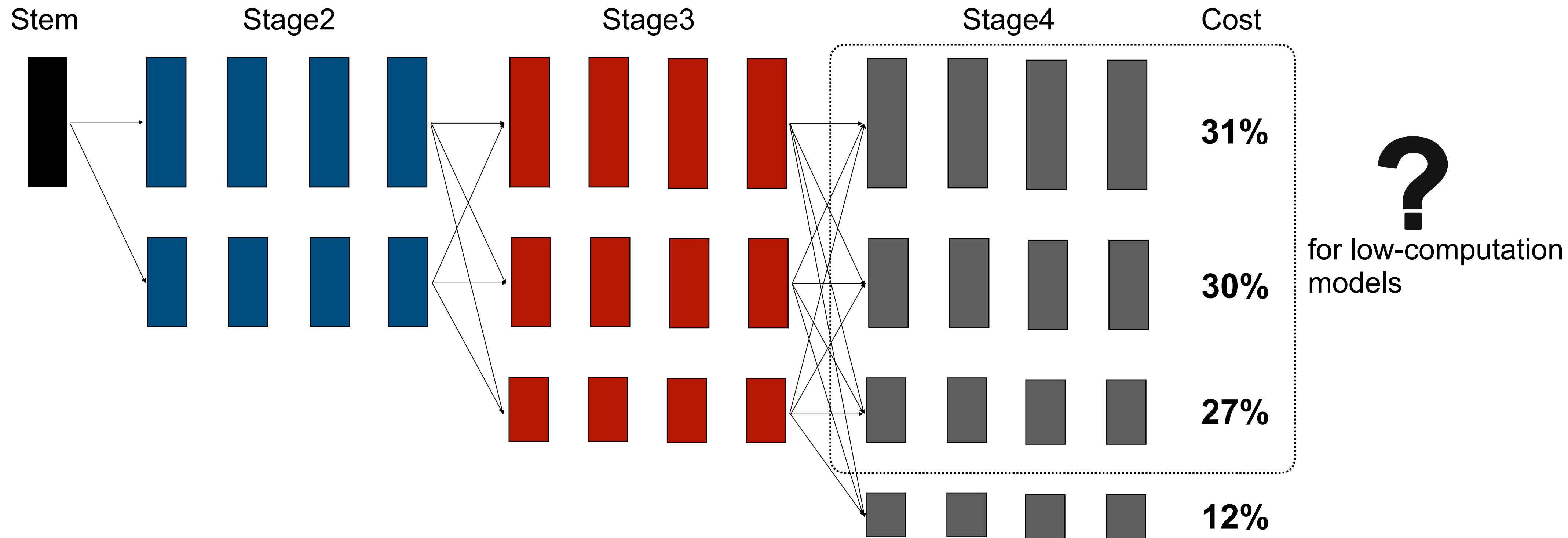
Most of the computational cost comes from high-resolution branches.

High-Resolution Branches are the Key Bottleneck



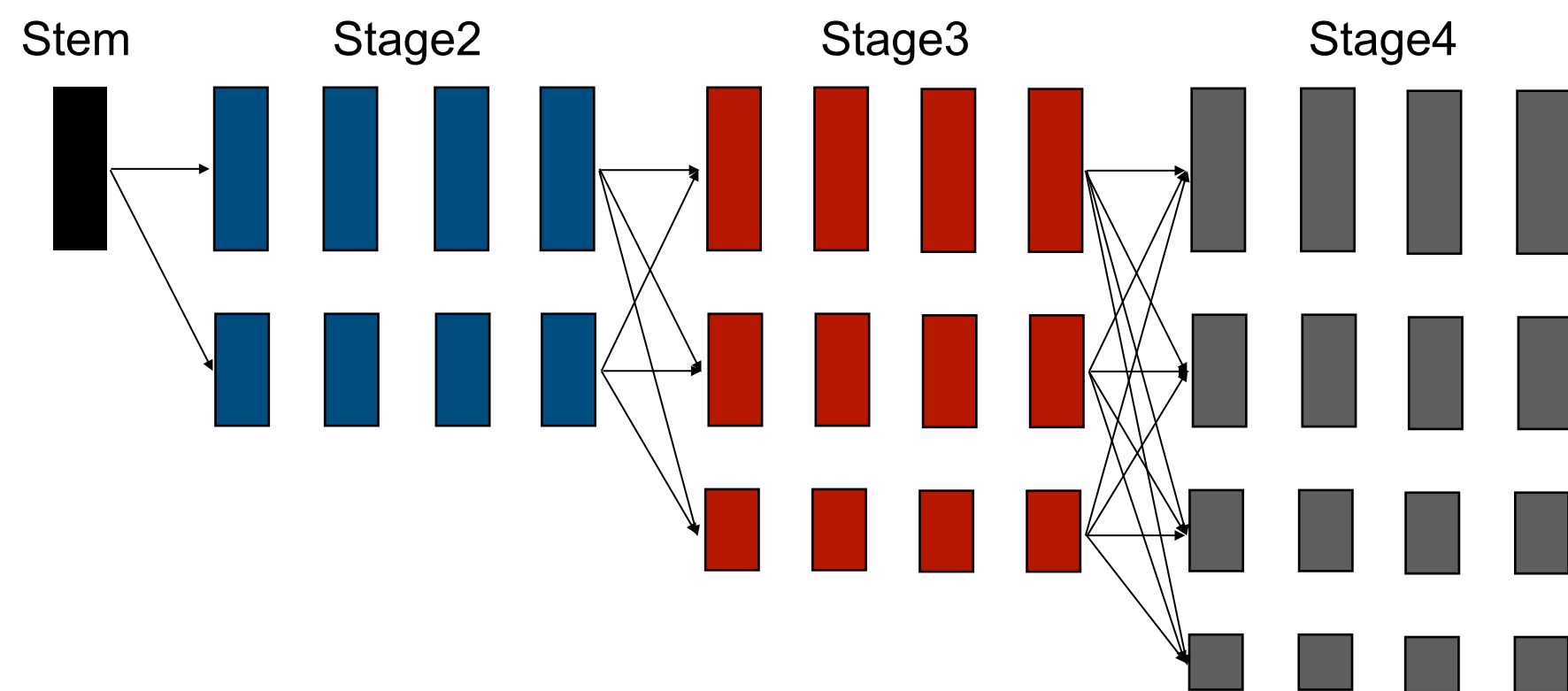
Most of the computational cost comes from high-resolution branches. Previous study in high-computation scenarios suggest that these high-resolution branches are essential.

High-Resolution Branches are the Key Bottleneck

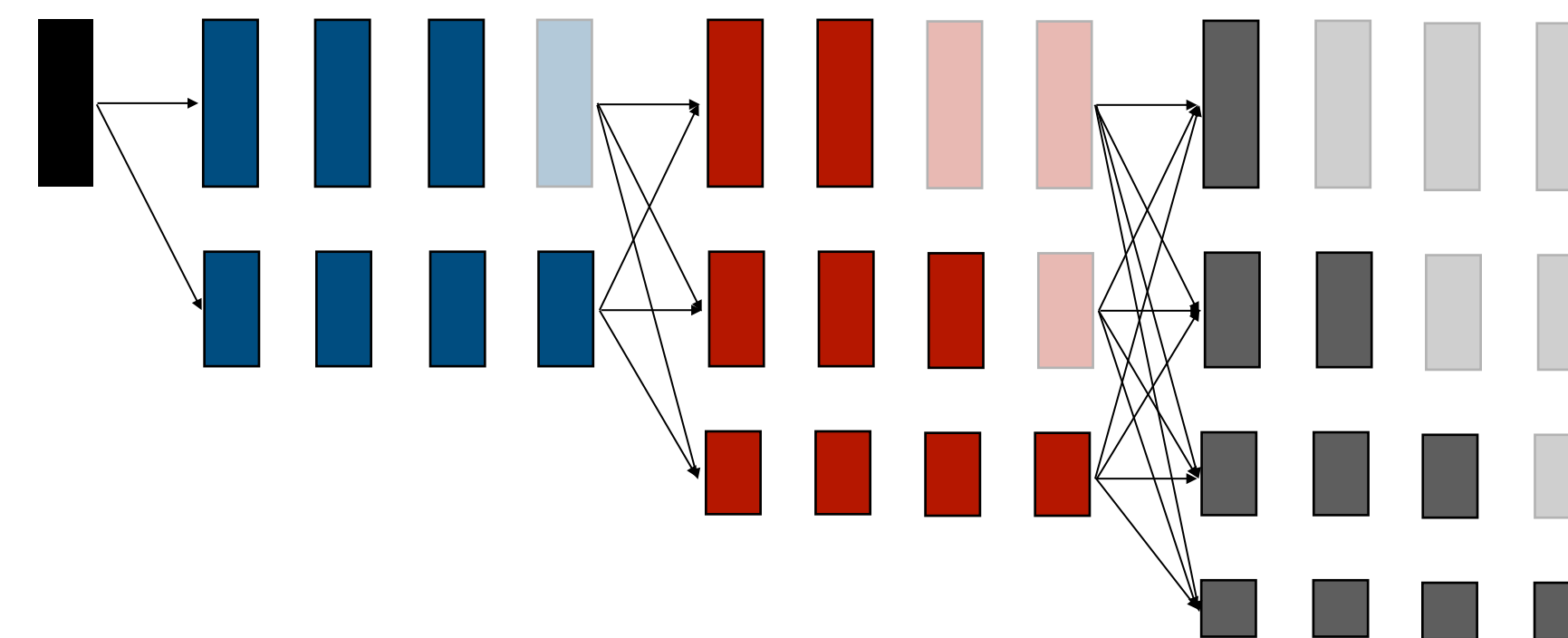


Most of the computational cost comes from high-resolution branches. Previous study in high-computation scenarios suggest that these high-resolution branches are essential.

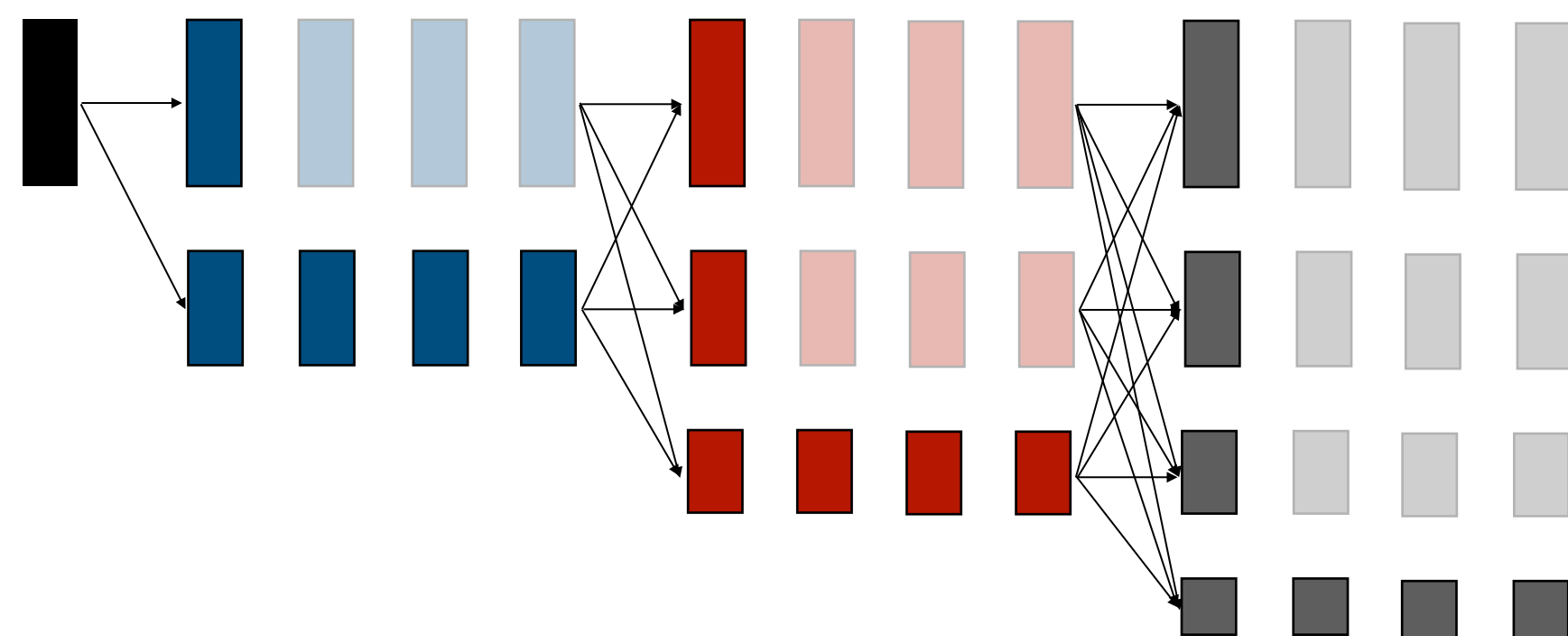
Gradual Shrinking Experiments



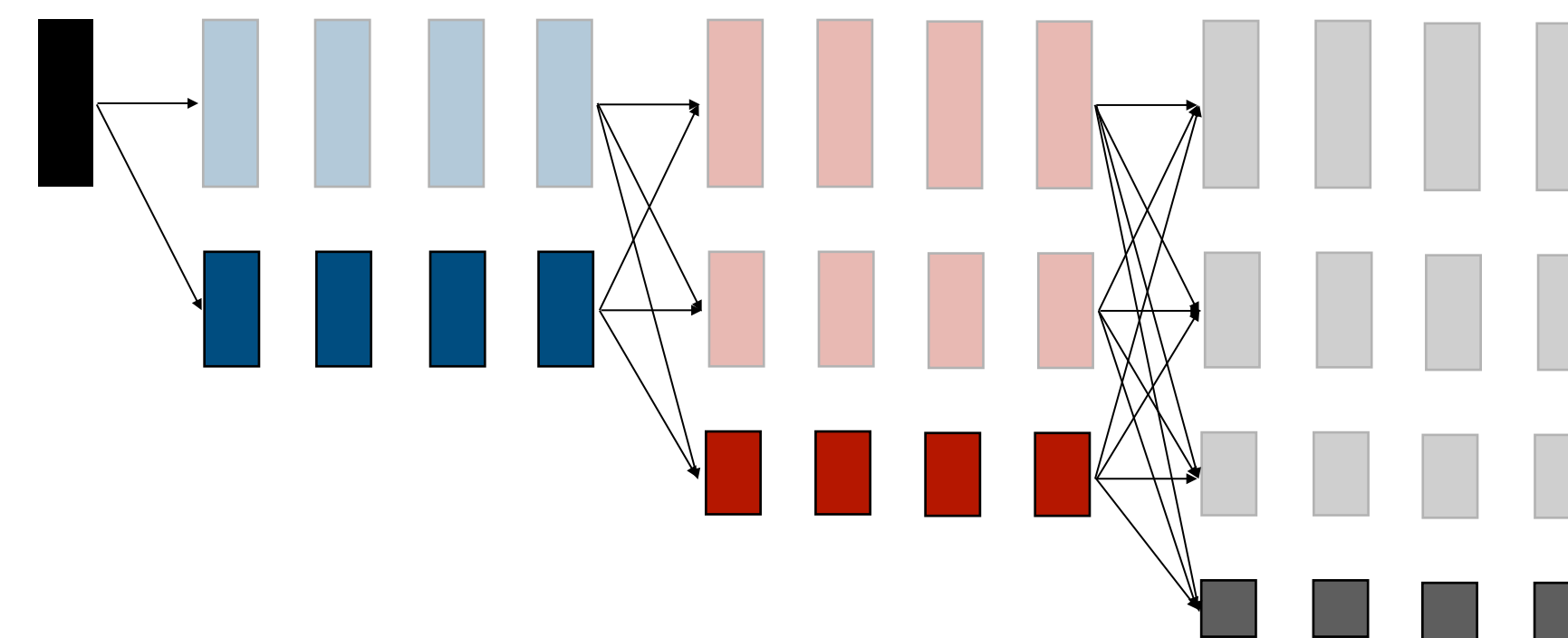
(a) HigherHRNet (Baseline), base channel = 16 (**Multi-Branch**)



(b) Shrink1, -29% blocks, base channel = 16



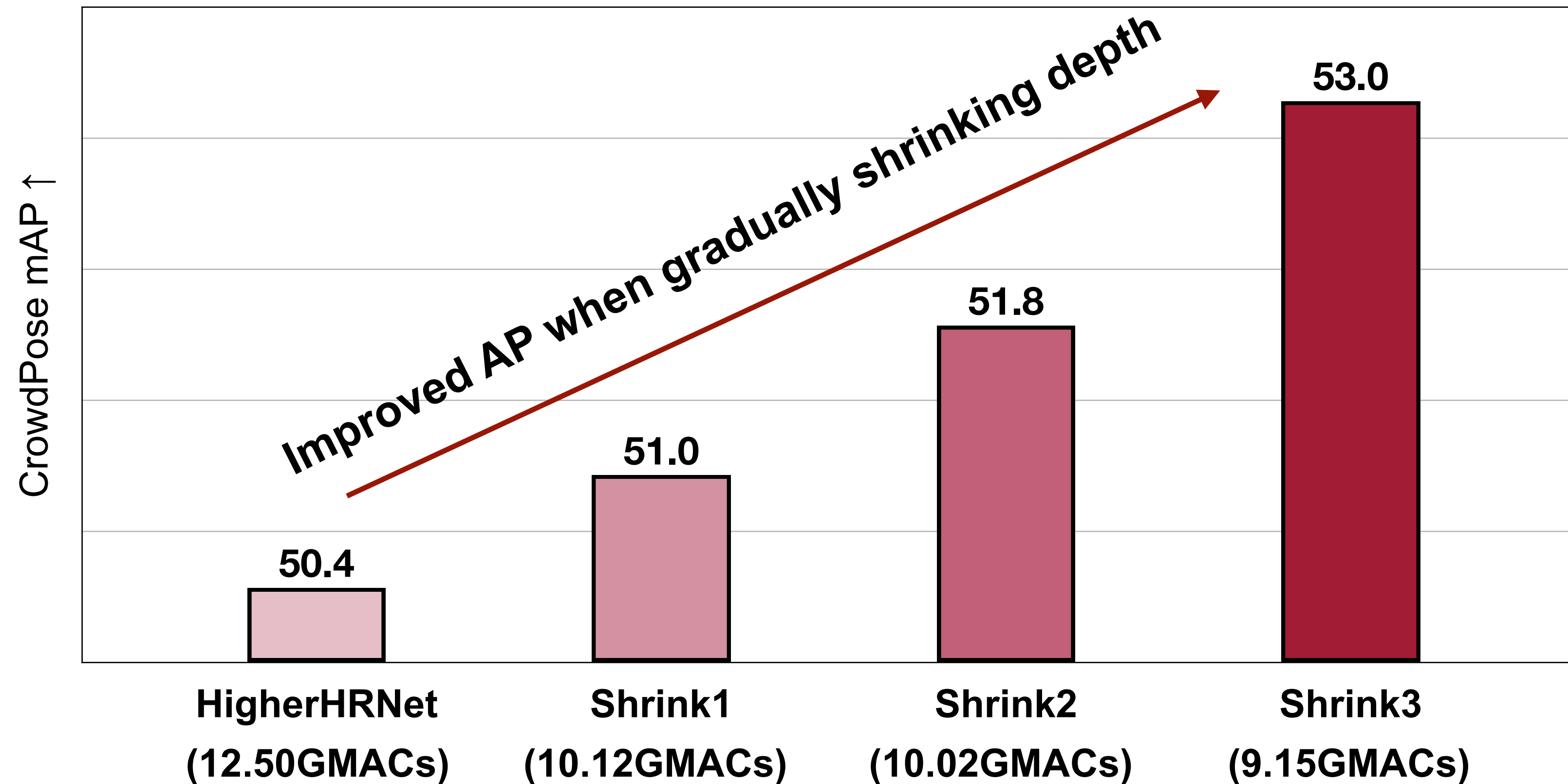
(c) Shrink2, -30% blocks, base channel = 18



(d) Shrink3, -33% blocks, base channel = 18 (**Single-Branch**)

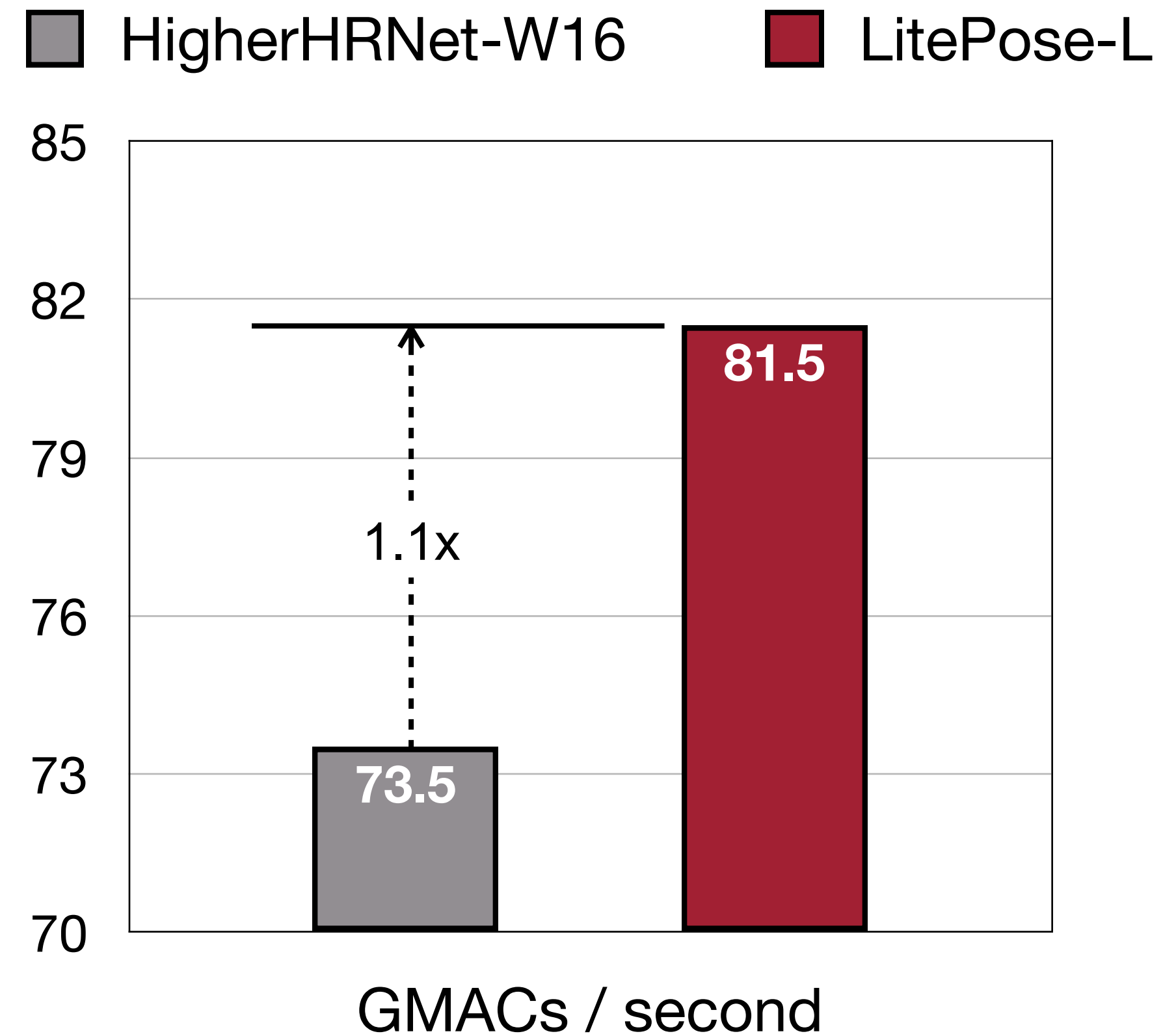
We gradually remove blocks in high-resolution branches starting from HigherHRNet. Removed blocks are shown in transparent.

Single Branch, Higher Performance



Removing high-resolution branches not only reduces the computational cost, but also improves the performance.

Single Branch, Higher Hardware Efficiency



Removing high-resolution branches makes the model more friendly for hardware, improving the GMACs / second by 1.1x.

Large Kernel Convolution is Important for Pose Estimation

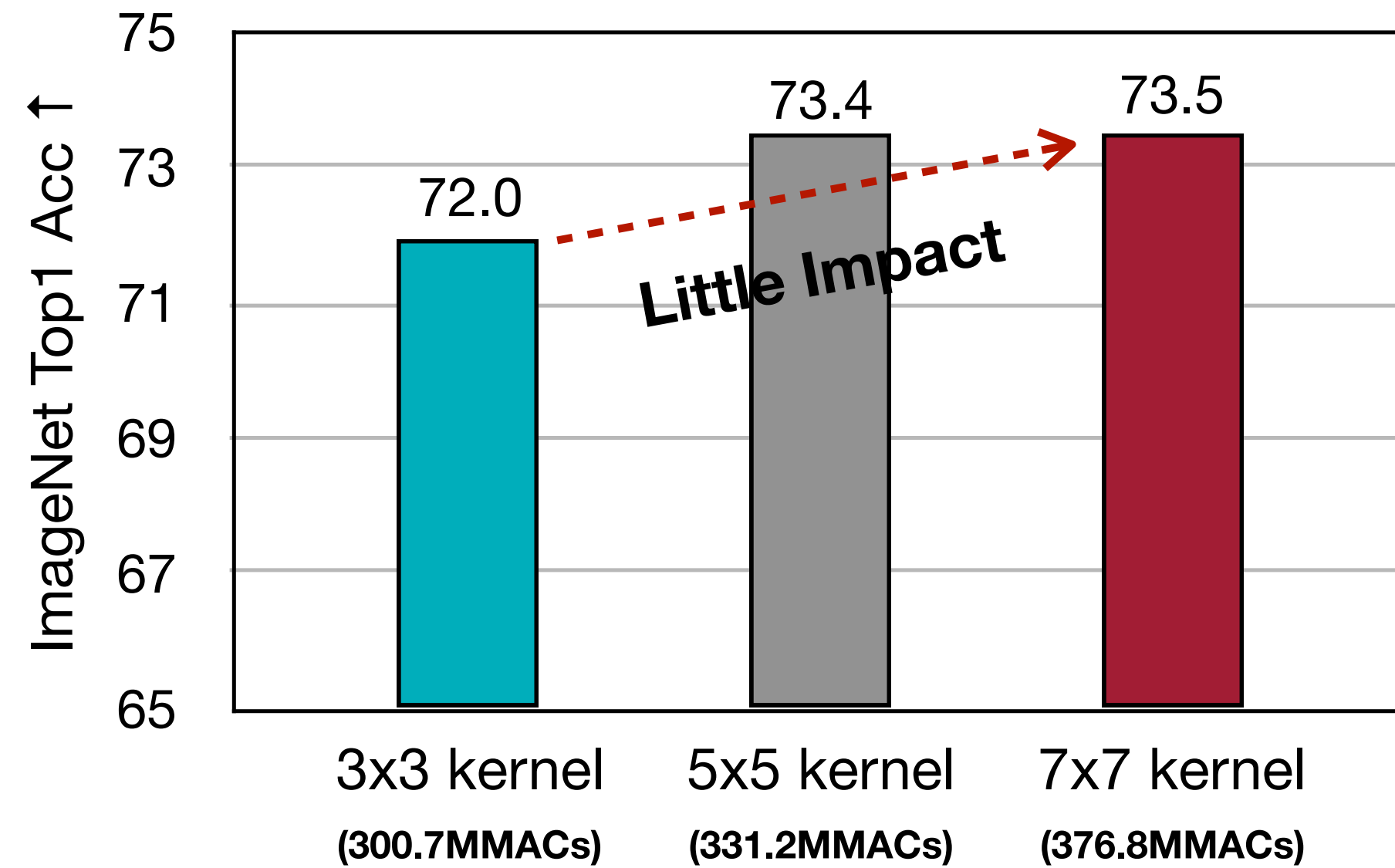
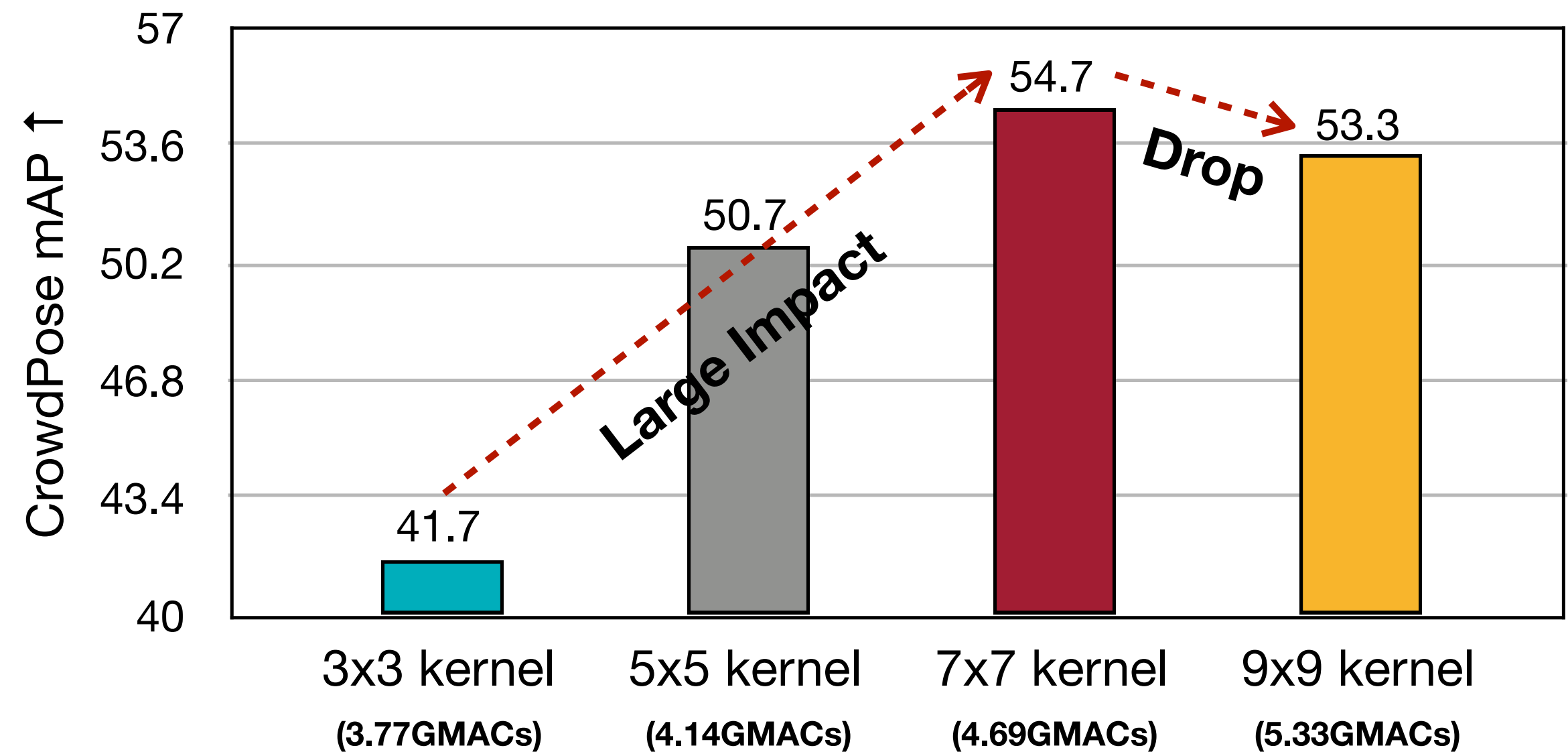


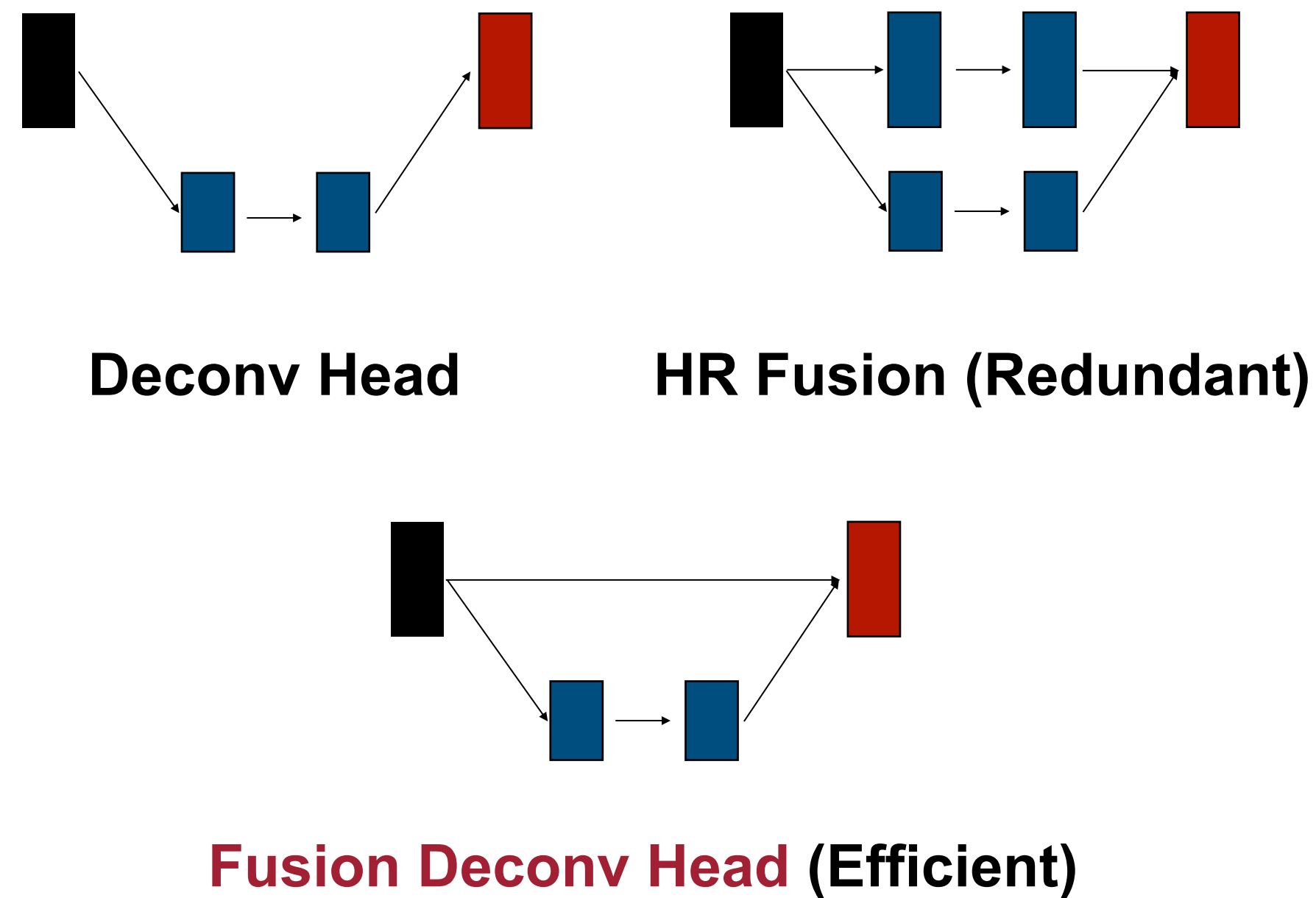
Image Classification



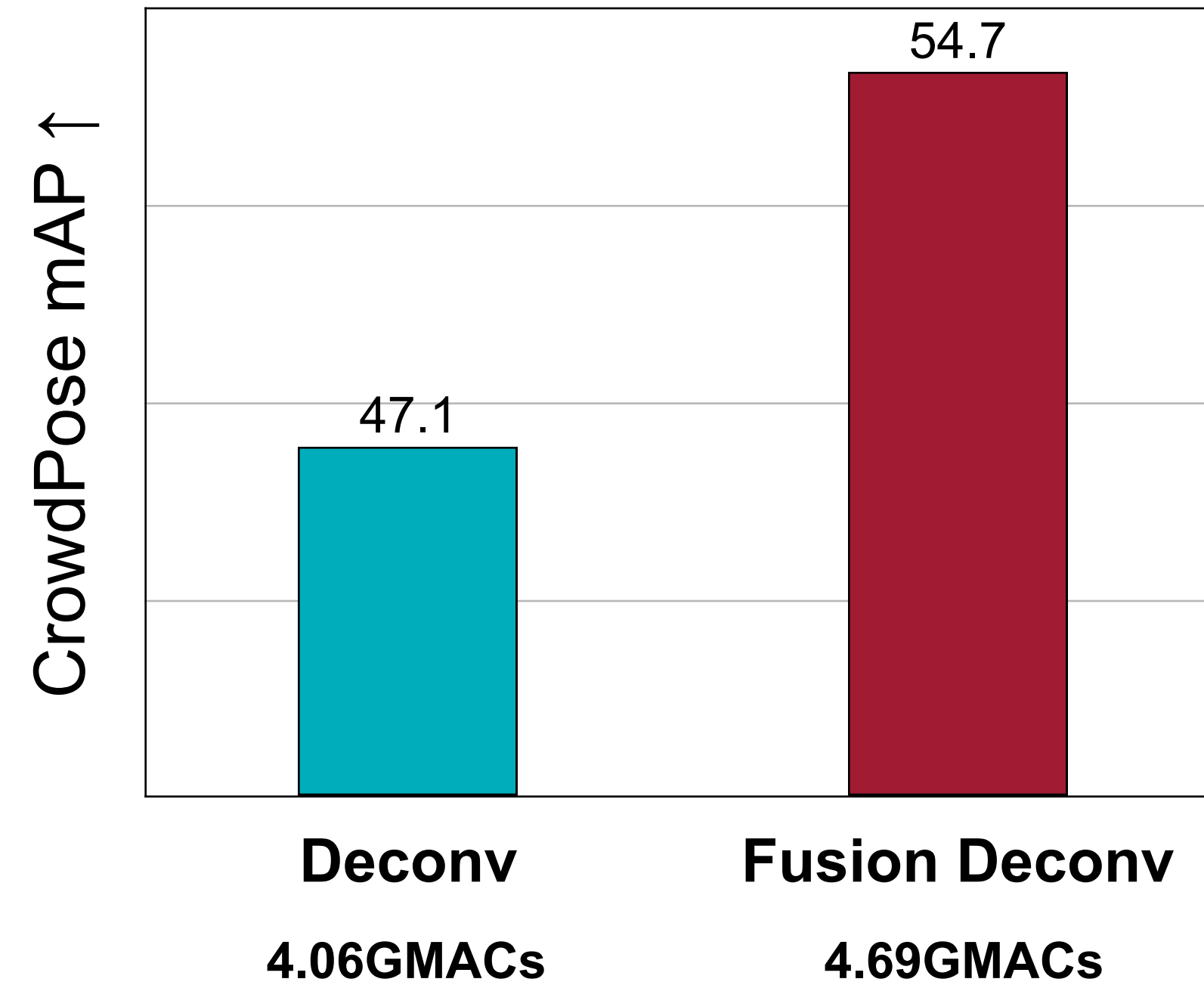
Human Pose Estimation

Unlike image classification, large kernel depthwise convolution plays a critical role in pose estimation. Increasing the kernel size from 3 to 7 improves the mAP by 13% on the CrowdPose dataset with little overhead.

Lightweight Fusion Deconv Head



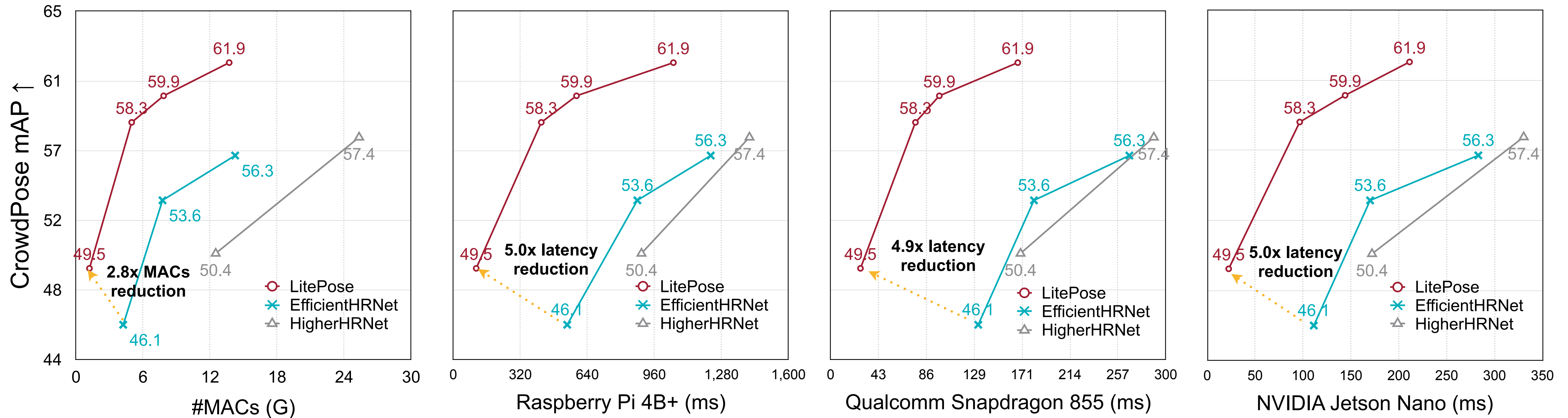
(a) Illustration of Heads



(b) Deconv vs. Fusion Deconv

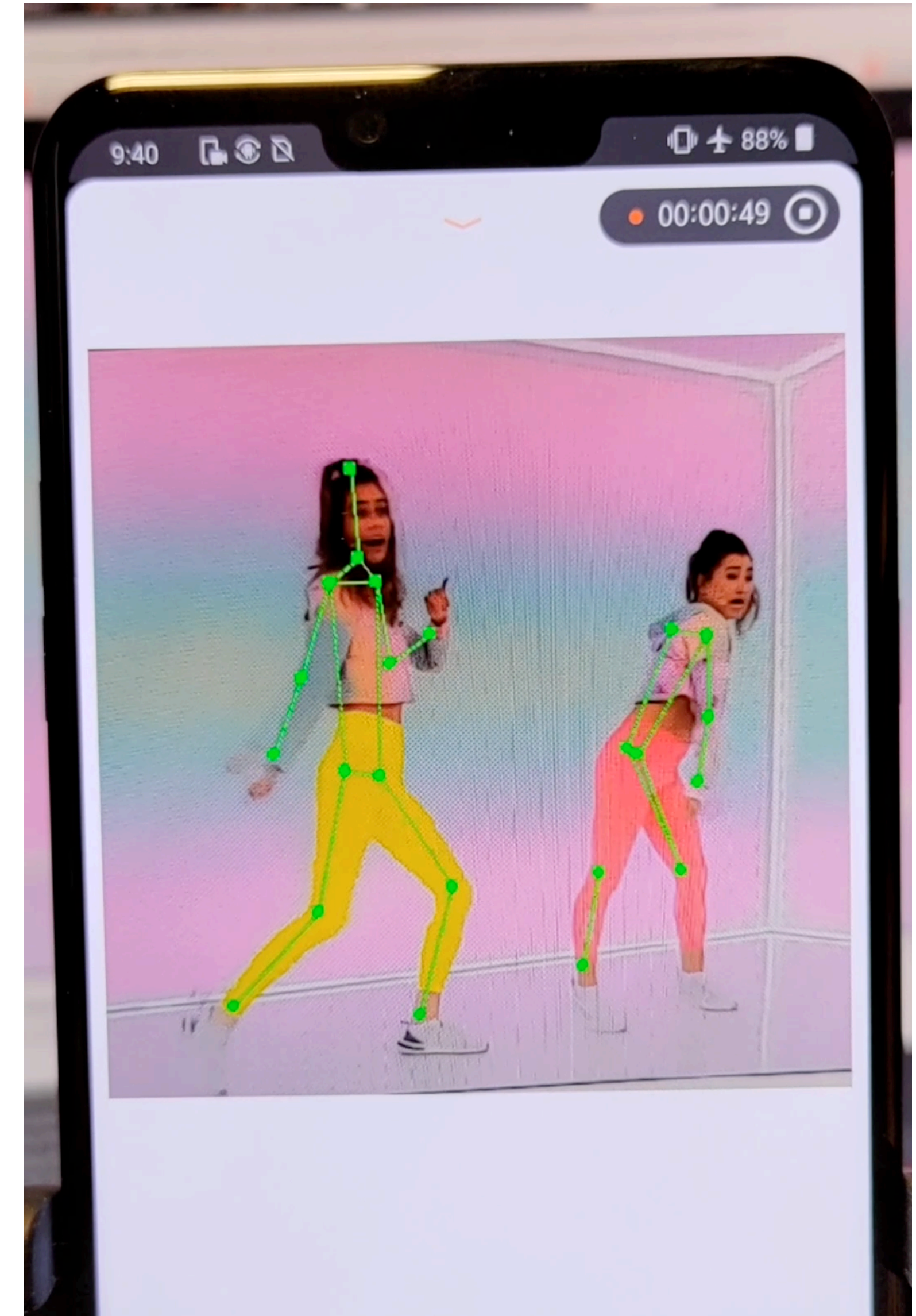
We employ the lightweight fusion deconv head to enable multi-resolution feature fusion without heavy high-resolution branches.

Compare with SOTA on the CrowdPose Dataset



2.8x MACs Reduction, 5.0x Speed Up

Real-Time Demo on LG
G8s ThinQ (Qualcomm
Snapdragon 855) with
LitePose-XS



Thank you!

<https://github.com/mit-han-lab/litepose>