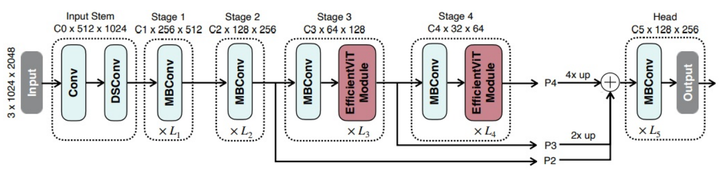
Table 2: **Detailed Architecture Configurations of Different EfficientViT Variants.** We build a series of models to fit different efficiency constraints. 'C' denotes the number of channels. 'L' denotes the number of blocks. 'H' is the height of the feature map, and 'W' is the width of the feature map.

Variants	Feature Map Shape	EfficientViT-B0	EfficientViT-B1	EfficientViT-B2	EfficientViT-B3
Input Stem	$C \times \frac{H}{2} \times \frac{W}{2}$	C = 8, L = 1	C = 16, L = 1	C = 24, L = 1	C = 32, L = 1
Stage1	$C \times \frac{H}{4} \times \frac{W}{4}$	C = 16, L = 2	C = 32, L = 2	C = 48, L = 3	C = 64, L = 4
Stage2	$C \times \frac{H}{8} \times \frac{W}{8}$	C = 32, L = 2	C = 64, L = 3	C = 96, L = 4	C = 128, L = 6
Stage3	$C \times \frac{H}{16} \times \frac{W}{16}$	C = 64, L = 2	C = 128, L = 3	C = 192, L = 4	C = 256, L = 6
Stage4	$C \times \frac{H}{32} \times \frac{W}{32}$	C = 128, L = 2	C = 256, L = 4	C = 384, L = 6	C = 512, L = 9
Head	$C \times \frac{H}{8} \times \frac{W}{8}$	C = 32, L = 1	C = 64, L = 3	C = 96, L = 3	C = 128, L = 3



- 降维和升维
- 跨通道信息交互
- 增加非线性特性

Figure 4: **Macro Architecture of EfficientViT.** We adopt the standard backbone-head/encoder-decoder design. In the backbone, we insert our lightweight MSA modules in Stages 3 and 4. Following the common practice, we feed the features from the last three stages (P2, P3, and P4) to the head. We use addition to fuse these features for simplicity and efficiency. As we already have lightweight MSA modules in the backbone, we adopt a simple head design that consists of several MBConv blocks and output layers.