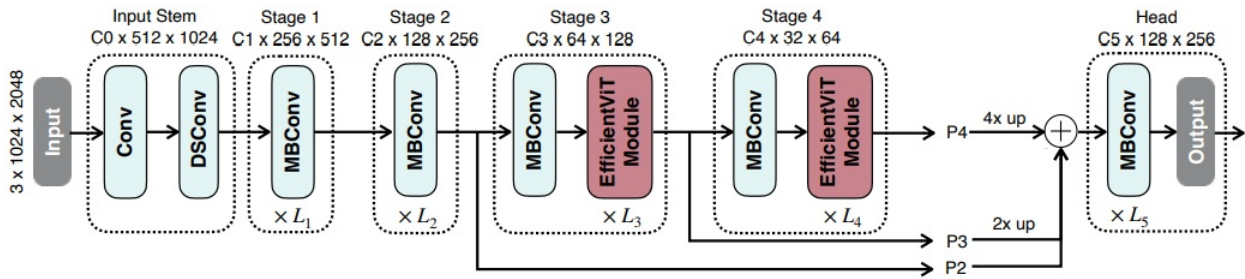


Table 2: **Detailed Architecture Configurations of Different EfficientViT Variants.** We build a series of models to fit different efficiency constraints. ‘C’ denotes the number of channels. ‘L’ denotes the number of blocks. ‘H’ is the height of the feature map, and ‘W’ is the width of the feature map.

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.

C: 3,244,244->32,112,112
C: 32,112,112->32,112,112

Conv_3x3_s2+bn+gelu

ResBlock
Conv_3x3+bn+gelu
Conv_3x3+bn
shortcut

stage0

C: 32->512->64,56,56
C: 64->256->64,56,56

FusedMBConv
Conv_3x3_s2+bn+gelu
Conv_1x1+bn

FusedMBConv
Conv_3x3+bn+gelu
Conv_1x1+bn
shortcut

stage1

C: 64->1024->128,28,28
C: 128->512->128,28,28

FusedMBConv
Conv_3x3_s2+bn+gelu
Conv_1x1+bn

FusedMBConv
Conv_3x3+bn+gelu
Conv_1x1+bn
shortcut

stage2

C: 128->2048->256,14,14
C: 256->1024->256,14,14

MBConv
Conv_1x1+gelu
Conv_3x3_s2+gelu
Conv_1x1+bn

MBConv
Conv_1x1+gelu
Conv_3x3+gelu
Conv_1x1+bn
shortcut

stage3

C: 256->6144->512,7,7
C: 512,7,7->512,7,7
C: 512,7,7->3072->512,7,7

MBConv
Conv_1x1+gelu
Conv_3x3_s2+gelu
Conv_1x1+bn

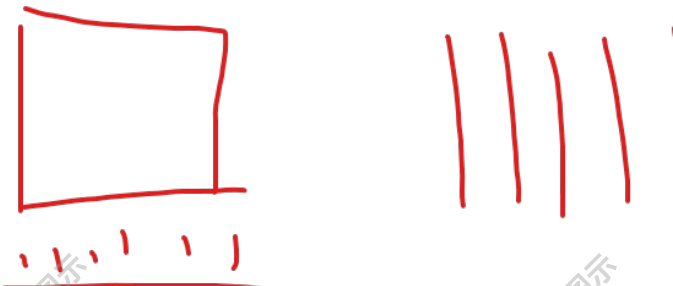
EfficientViTBlock x 6
multi-scale-attention
qkv: conv_1x1
aggreg: conv_5x5/1x1
kernel_fun: Relu
proj: conv1x1+bn
MBConv
Conv_1x1+gelu
Conv_3x3+gelu
Conv_1x1+bn

stage4

C: 512->3072
C: 3072->3200
C: 3200->1000

MBConv
Conv_1x1+bn+gelu
AdaptiveAvgPool
Linear+ln+gelu
Linear

Head


$$\text{Sim}(K) = \exp\left(\frac{QK^T}{\sqrt{d}}\right)$$

Handwritten notes: \checkmark 32x32, 32x32