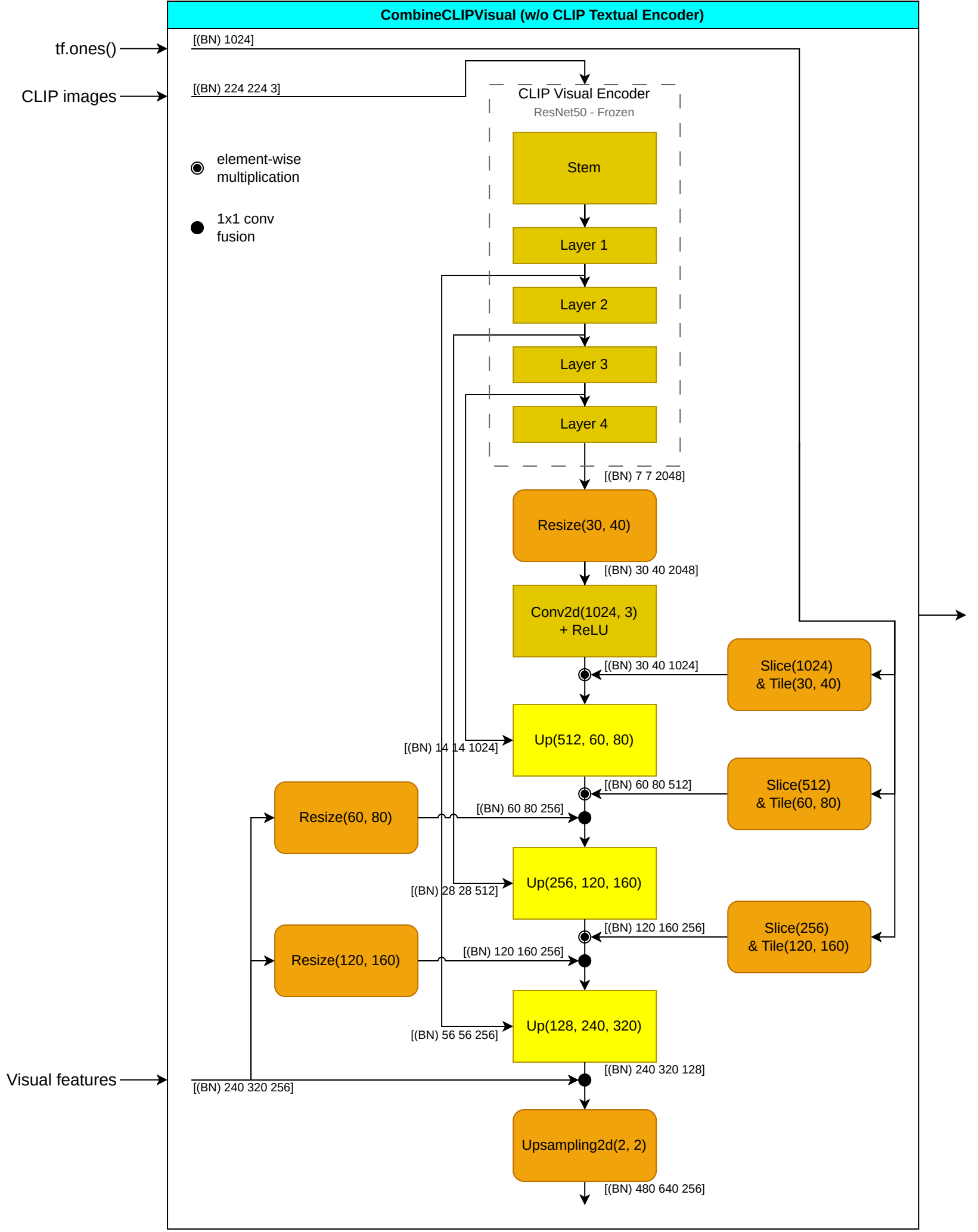
CombineCLIPVisual (w/o CLIP Textual Encoder)
tf.ones()
CLIP images
Visual features
[(BN) 1024]
[(BN) 224 224 3]
[(BN) 240 320 256]
CLIP Visual Encoder
ResNet50 - Frozen
Stem
Layer 1
Layer 2
Layer 3
Layer 4
element-wise multiplication
1x1 conv fusion
[(BN) 7 7 2048]
Resize(30, 40)
[(BN) 30 40 2048]
Conv2d(1024, 3) + ReLU
[(BN) 30 40 1024]
Slice(1024) & Tile(30, 40)
[(BN) 14 14 1024]
Up(512, 60, 80)
[(BN) 60 80 512]
Slice(512) & Tile(60, 80)
Resize(60, 80)
[(BN) 60 80 256]
[(BN) 28 28 512]
Up(256, 120, 160)
[(BN) 120 160 256]
Slice(256) & Tile(120, 160)
Resize(120, 160)
[(BN) 120 160 256]
[(BN) 56 56 256]
Up(128, 240, 320)
[(BN) 240 320 128]
Upsampling2d(2, 2)
[(BN) 480 640 256]