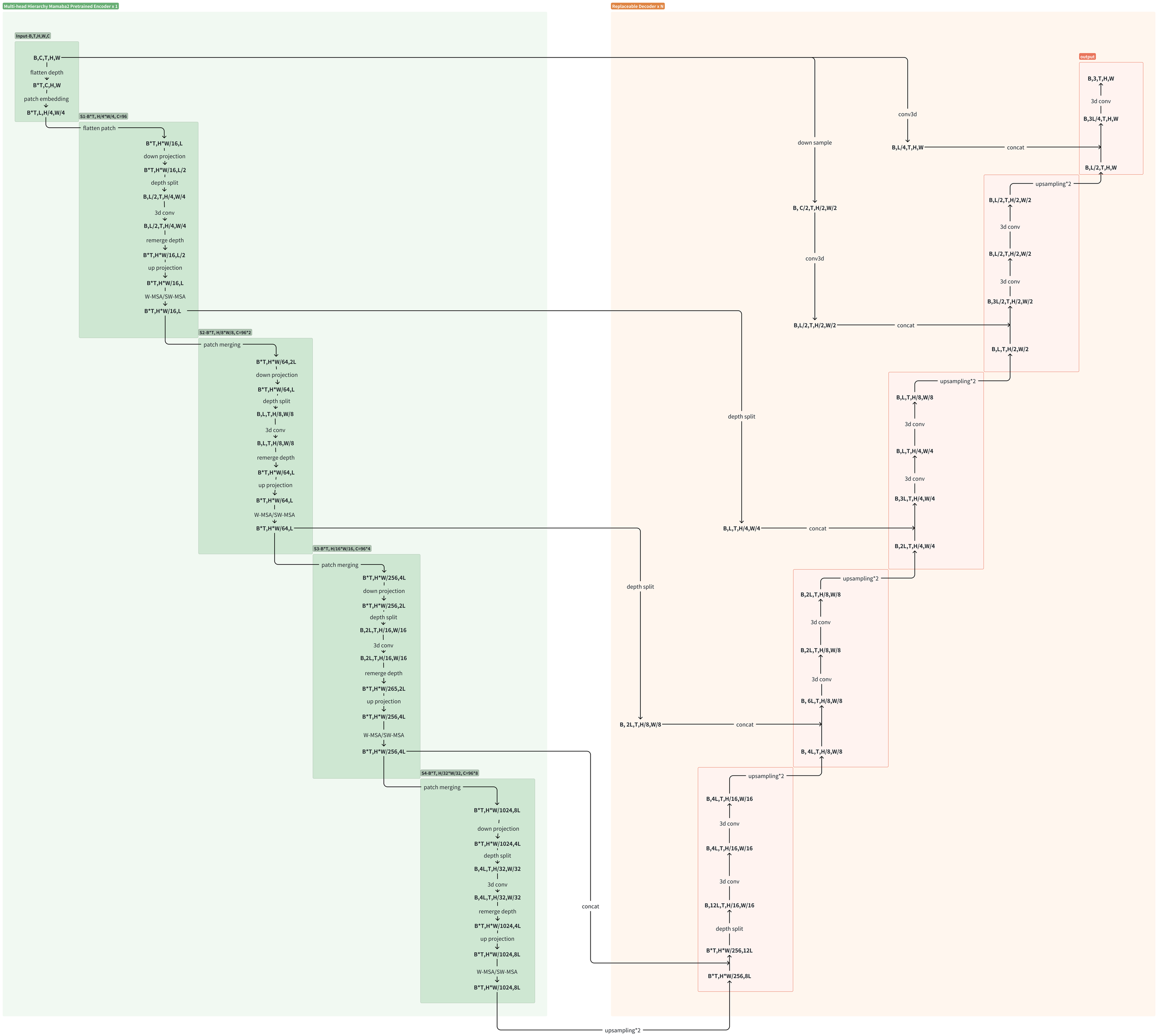
Multi-head Hierarchy Mamaba2 Pretrained Encoder x 1
Replaceable Decoder x N
Input-B,T,H,W,C
B,C,T,H,W
flatten depth
B*T,C,H,W
patch embedding
B*T,L,H/4,W/4
S1-B*T, H/4*W/4, C=96
flatten patch
B*T,H*W/16,L
down projection
B*T,H*W/16,L/2
depth split
B,L/2,T,H/4,W/4
3d conv
B,L/2,T,H/4,W/4
remerge depth
B*T,H*W/16,L/2
up projection
B*T,H*W/16,L
W-MSA/SW-MSA
B*T,H*W/16,L
S2-B*T, H/8*W/8, C=96*2
patch merging
B*T,H*W/64,2L
down projection
B*T,H*W/64,L
depth split
B,L,T,H/8,W/8
3d conv
B,L,T,H/8,W/8
remerge depth
B*T,H*W/64,L
up projection
B*T,H*W/64,L
W-MSA/SW-MSA
B*T,H*W/64,L
S3-B*T, H/16*W/16, C=96*4
patch merging
B*T,H*W/256,4L
down projection
B*T,H*W/256,2L
depth split
B,2L,T,H/16,W/16
3d conv
B,2L,T,H/16,W/16
remerge depth
B*T,H*W/265,2L
up projection
B*T,H*W/256,4L
W-MSA/SW-MSA
B*T,H*W/256,4L
S4-B*T, H/32*W/32, C=96*8
patch merging
B*T,H*W/1024,8L
down projection
B*T,H*W/1024,4L
depth split
B,4L,T,H/32,W/32
3d conv
B,4L,T,H/32,W/32
remerge depth
B*T,H*W/1024,4L
up projection
B*T,H*W/1024,8L
W-MSA/SW-MSA
B*T,H*W/1024,8L
upsampling*2
down sample
B, C/2,T,H/2,W/2
conv3d
B,L/2,T,H/2,W/2
concat
B,L,T,H/2,W/2
depth split
B,L,T,H/4,W/4
concat
B,2L,T,H/4,W/4
depth split
B, 2L,T,H/8,W/8
concat
B, 4L,T,H/8,W/8
concat
B,12L,T,H/16,W/16
B*T,H*W/256,8L
upsampling*2
B,4L,T,H/16,W/16
3d conv
B,4L,T,H/16,W/16
3d conv
B,4L,T,H/16,W/16
upsampling*2
B,2L,T,H/8,W/8
3d conv
B,2L,T,H/8,W/8
3d conv
B, 6L,T,H/8,W/8
upsampling*2
B,L,T,H/8,W/8
3d conv
B,L,T,H/4,W/4
3d conv
B,3L,T,H/4,W/4
upsampling*2
B,L/2,T,H/2,W/2
3d conv
B,L/2,T,H/2,W/2
3d conv
B,3L/2,T,H/2,W/2
output
B,3,T,H,W
3d conv
B,3L/4,T,H,W
B,L/2,T,H,W
conv3d
B,L/4,T,H,W
concat