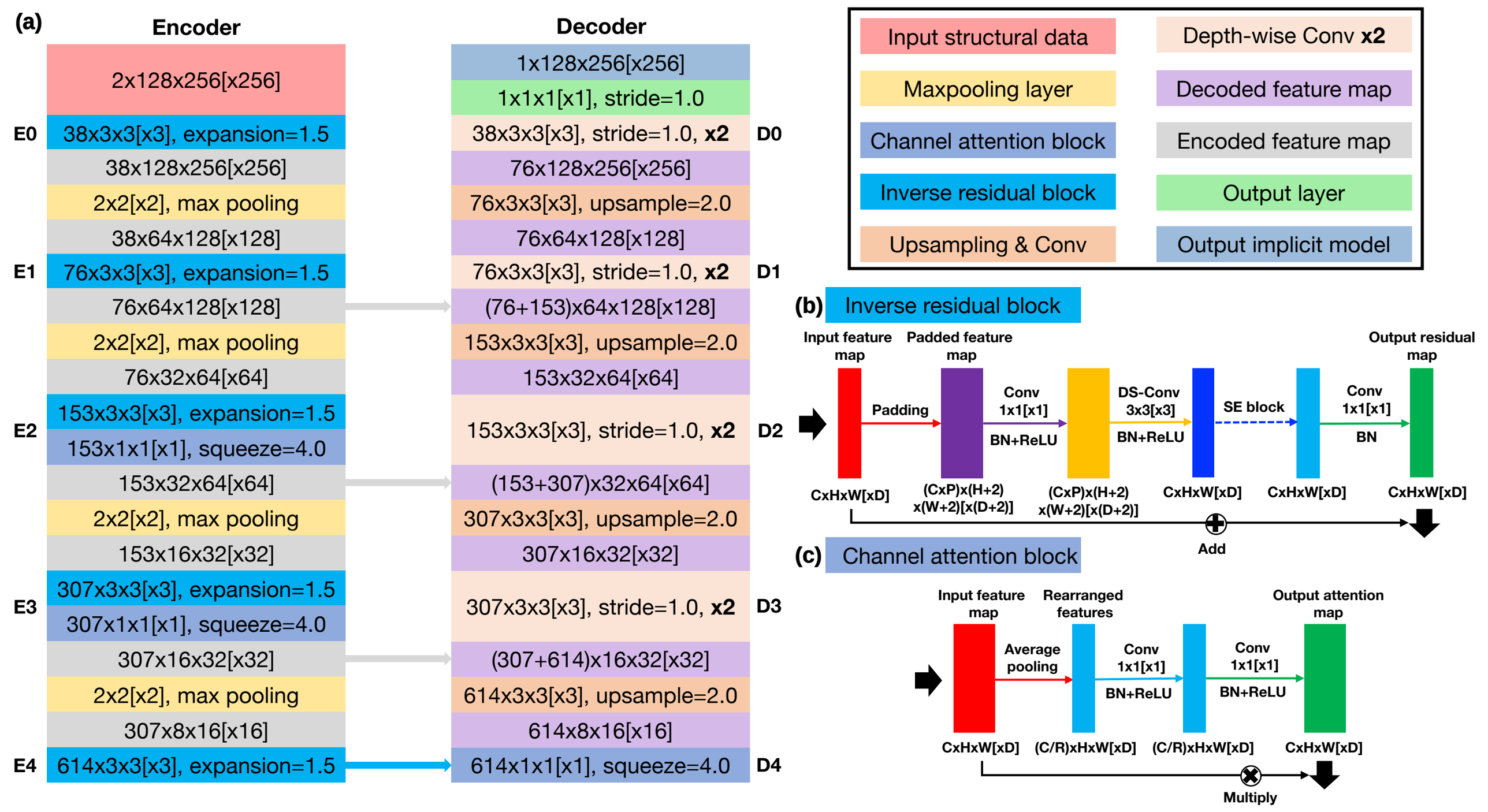
(a)
Encoder
Decoder
2x128x256[x256]
1x128x256[x256]
1x1x1[x1], stride=1.0
E0
38x3x3[x3], expansion=1.5
38x3x3[x3], stride=1.0, x2
D0
38x128x256[x256]
76x128x256[x256]
2x2[x2], max pooling
76x3x3[x3], upsample=2.0
38x64x128[x128]
76x64x128[x128]
E1
76x3x3[x3], expansion=1.5
76x3x3[x3], stride=1.0, x2
D1
76x64x128[x128]
(76+153)x64x128[x128]
2x2[x2], max pooling
153x3x3[x3], upsample=2.0
76x32x64[x64]
153x32x64[x64]
E2
153x3x3[x3], expansion=1.5
153x1x1[x1], squeeze=4.0
153x3x3[x3], stride=1.0, x2
D2
153x32x64[x64]
(153+307)x32x64[x64]
2x2[x2], max pooling
307x3x3[x3], upsample=2.0
153x16x32[x32]
307x16x32[x32]
E3
307x3x3[x3], expansion=1.5
307x1x1[x1], squeeze=4.0
307x3x3[x3], stride=1.0, x2
D3
307x16x32[x32]
(307+614)x16x32[x32]
2x2[x2], max pooling
614x3x3[x3], upsample=2.0
307x8x16[x16]
614x8x16[x16]
E4
614x3x3[x3], expansion=1.5
614x1x1[x1], squeeze=4.0
D4
Input structural data
Depth-wise Conv x2
Maxpooling layer
Decoded feature map
Channel attention block
Encoded feature map
Inverse residual block
Output layer
Upsampling & Conv
Output implicit model
(b) Inverse residual block
Input feature map
Padded feature map
Output residual map
Padding
Conv 1x1[x1]
BN+ReLU
DS-Conv 3x3[x3]
BN+ReLU
SE block
Conv 1x1[x1]
BN
CxHxW[xD]
(CxP)x(H+2) x(W+2)[x(D+2)]
(CxP)x(H+2) x(W+2)[x(D+2)]
CxHxW[xD]
CxHxW[xD]
CxHxW[xD]
Add
(c) Channel attention block
Input feature map
Rearranged features
Output attention map
Average pooling
Conv 1x1[x1]
BN+ReLU
Conv 1x1[x1]
BN+ReLU
CxHxW[xD]
(C/R)xHxW[xD]
(C/R)xHxW[xD]
CxHxW[xD]
Multiply