ASPP

Image Pooling

Conv 3 × 3 rate 6

Conv 3 × 3 rate 12

Conv 3 × 3 rate 18

Conv 1 × 1

MS-Encoder

Decoder

Reshape

Linear Projection

Transformer Layer

(n = 12)

Transformer Layer

MLP

Layer Norm

Multi-Head Attention

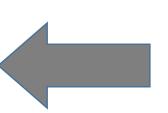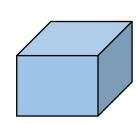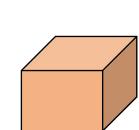Layer Norm

SA-Encoder

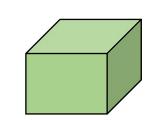Downsample

Upsample

Concatenate

Element-wise Addition

SA-Encoder Block

MS-Encoder Block

Decoder Block

Skip-Connection