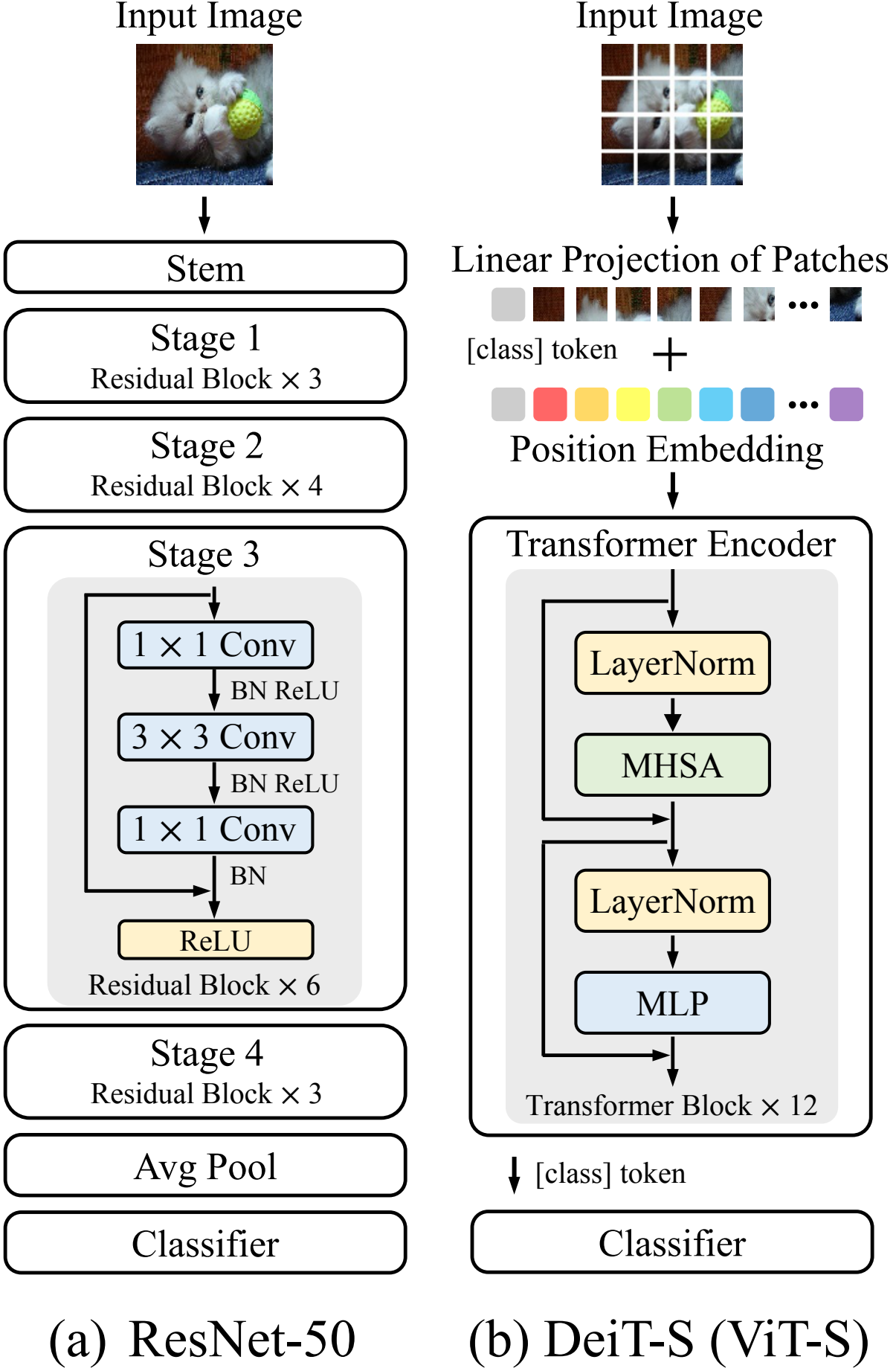
Input Image
Stem
Stage 1
Residual Block × 3
Stage 2
Residual Block × 4
Stage 3
1 × 1 Conv
BN ReLU
3 × 3 Conv
BN ReLU
1 × 1 Conv
BN
ReLU
Residual Block × 6
Stage 4
Residual Block × 3
Avg Pool
Classifier
(a) ResNet-50
Input Image
Linear Projection of Patches
[class] token
Position Embedding
Transformer Encoder
LayerNorm
MHSA
LayerNorm
MLP
Transformer Block × 12
[class] token
Classifier
(b) DeiT-S (ViT-S)