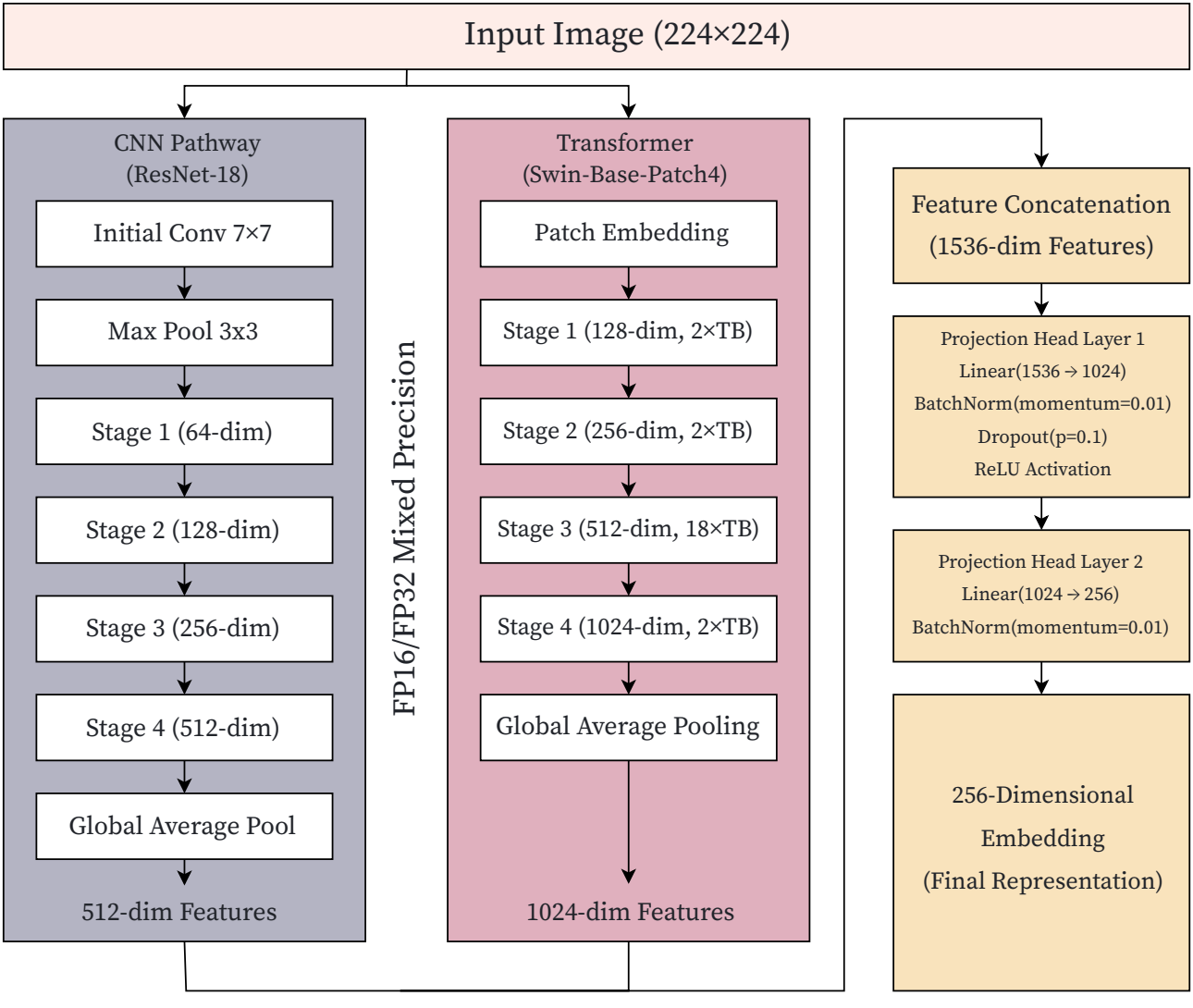
Input Image (224×224)
CNN Pathway
(ResNet-18)
Initial Conv 7×7
Max Pool 3x3
Stage 1 (64-dim)
Stage 2 (128-dim)
Stage 3 (256-dim)
Stage 4 (512-dim)
Global Average Pool
512-dim Features
FP16/FP32 Mixed Precision
Transformer
(Swin-Base-Patch4)
Patch Embedding
Stage 1 (128-dim, 2×TB)
Stage 2 (256-dim, 2×TB)
Stage 3 (512-dim, 18×TB)
Stage 4 (1024-dim, 2×TB)
Global Average Pooling
1024-dim Features
Feature Concatenation
(1536-dim Features)
Projection Head Layer 1
Linear(1536 → 1024)
BatchNorm(momentum=0.01)
Dropout(p=0.1)
ReLU Activation
Projection Head Layer 2
Linear(1024 → 256)
BatchNorm(momentum=0.01)
256-Dimensional
Embedding
(Final Representation)