CNN U-Net vs Transformer Encoder–Decoder (Binary Segmentation)

Stage	U-Net Layer	Feature Dim (U-Net)	Role	Transformer Layer	Feature Dim (Transformer)	Role
Input	Raw Image	256×256×3	Input RGB image	Image split into 16×16 patches	256 tokens × 768 dims	Convert image into tokens
Encoder – Stage 1	Conv(3×3,64) → Conv(3×3,64), MaxPool	128×128×64	Extract low-level edges/textures	Linear patch embedding	256×768	Represent patches as embeddings
Encoder – Stage 2	Conv(128) + Pool	64×64×128	Mid-level features	Transformer encoder block	256×768	Global context via self-attention
Encoder – Stage 3	Conv(256) + Pool	32×32×256	Higher-level structures	Deeper Transformer encoders	256×768	Capture long-range dependencies
Encoder – Stage 4	Conv(512) + Pool	16×16×512	Abstract semantics	Transformer latent tokens	256×768	Global semantic representation
Bottleneck	Conv(1024)	16×16×1024	Compact latent features	Transformer token sequence	256×768	Final latent representation
Skip Connections	Feature maps copied to decoder	128×128, 64×64, 32×32, 16×16	Preserve spatial detail	Token–feature fusion	Mixed (tokens + feature maps)	Inject local detail back
Decoder – Stage 1	UpConv(512) + Concat(encoder)	32×32×512	Reconstruct with skip info	Decoder cross-attention	Reshape tokens → 32×32×C	Decode tokens into spatial map
Decoder – Stage 2	UpConv(256) + Concat	64×64×256	Refine spatial resolution	Upsampling + projection	64×64×C	Reconstruct mid-level features
Decoder – Stage 3	UpConv(128) + Concat	128×128×128	Sharpen object boundaries	Progressive reconstruction	128×128×C	Refine mask details
Decoder – Final	UpConv(64) + Concat → Conv(1×1,1)	256×256×1	Output binary mask	Projection head + Sigmoid	256×256×1	Output binary mask
Feature Extraction	Convolutions with local receptive fields	Hierarchical spatial maps	Local texture + shape features	Self-attention with global receptive field	Token sequence preserved	Global dependencies captured
Mask Generation	Upsampling with transposed convs	Pixel-level spatial maps	Reconstruct detailed mask	Linear projection of decoded tokens	Pixel grid restored	Reconstruct binary mask