Table 1. We conducted an ablation study with various arrangements to assess the impacts of MSA projection, dilation and attention aggregation in SWC, and the use of attention mixing on segmentation performance over the Synapse-multi organ dataset. To do this, we tested attention-mixing with both yes and no options, linear or convolutional MSA projection, dilations of 1,2 or 2,3, and attention-aggregation with either summation or concatenation. We performed 16 experiments by combining all possible configurations and found that the optimal setup (highlighted in bold) with the highest performance includes attention mixing, convolutional projected MSA, dilation of 2,3, and concatenation of attention in the SWC block.

Attention-mixing	MSA Projection	Dilations	Attention-aggregation	Mean	
				DICE	HD95
Yes	Linear	1,2	Summation	83.64	14.79
Yes	Linear	1,2	Concatenation	84.98	14.08
Yes	Linear	2,3	Summation	85.03	13.48
Yes	Linear	2,3	Concatenation	84.87	13.15
Yes	Convolution	1,2	Summation	83.32	15.01
Yes	Convolution	1,2	Concatenation	86.23	12.65
Yes	Convolution	2,3	Summation	85.43	12.01
Yes	**Convolution**	**2,3**	**Concatenation**	**86.95**	**11.08**
No	Linear	1,2	Summation	83.45	15.28
No	Linear	1,2	Concatenation	84.06	14.87
No	Linear	2,3	Summation	84.32	15.34
No	Linear	2,3	Concatenation	85.09	13.97
No	Convolution	1,2	Summation	83.67	16.54
No	Convolution	1,2	Concatenation	84.32	14.74
No	Convolution	2,3	Summation	83.69	15.65
No	Convolution	2,3	Concatenation	83.81	14.11

Table 2. Evaluation of Interpolation Effects on Performance (with Optimal Settings in other aspects): Bilinear interpolation demonstrates the highest Dice score and the lowest HD95 among the different interpolation methods.

Interpolations	Mean	
	DICE	HD95
Bilinear	**86.95**	**11.08**
Area	84.32	15.63
Bicubic	84.60	14.31
Nearest-exact	84.13	14.87

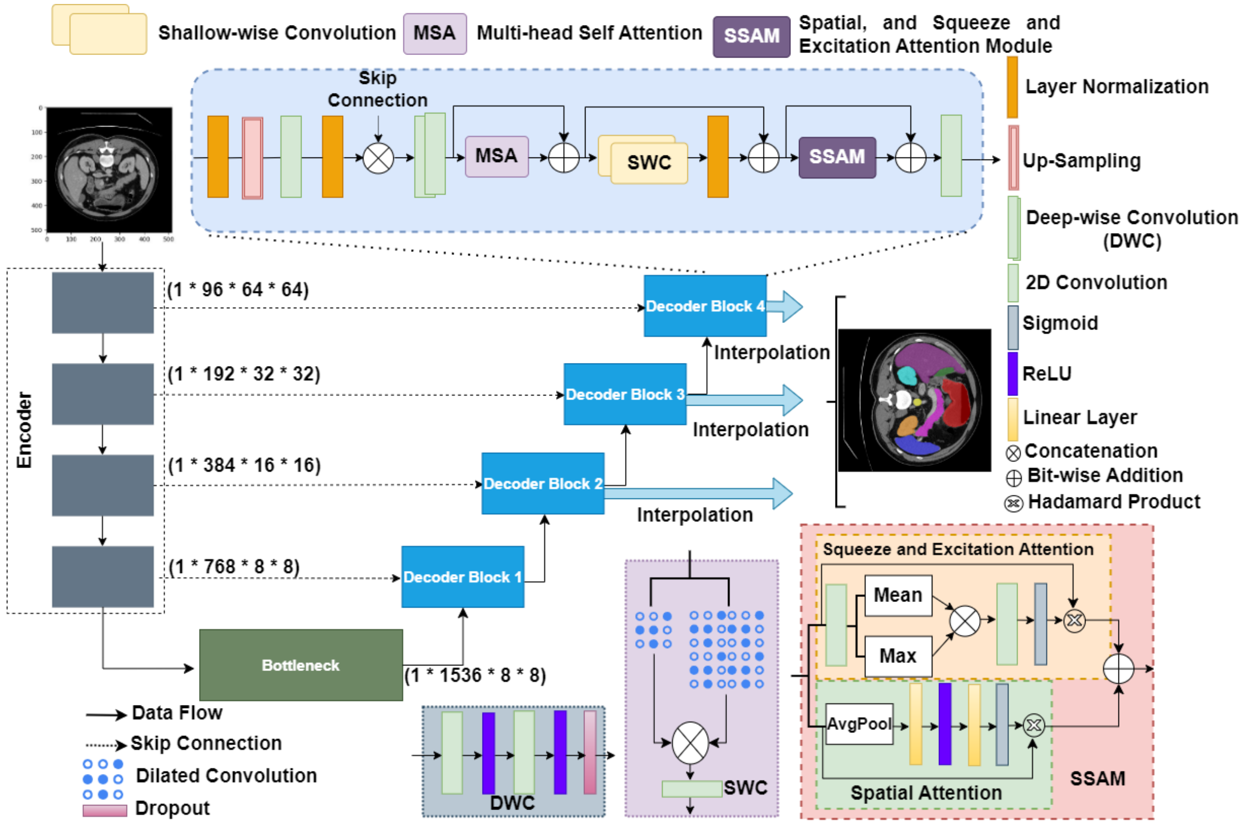

Figure 1. Medical Image Segmentation Transformer (MIST): **The figure shows the decoder blocks without mixing attentions**. The left side represents the encoder, utilizing a pre-trained MaxViT model. On the right side, the decoder generates segmentation maps. Each decoder block incorporates a convolutional projected MSA (Multi-head Self-Attention) to reduce computational cost and capture salient features. Additionally, depth-wise (deep and shallow) convolutions (DWC and SWC) are incorporated to extract relevant semantic features and enhance the kernel's receptive field, facilitating better long-range dependency.