Supplementary material of the paper: MetaUNETR: Rethinking Token Mixer Encoding for Efficient Multi-Organ Segmentation

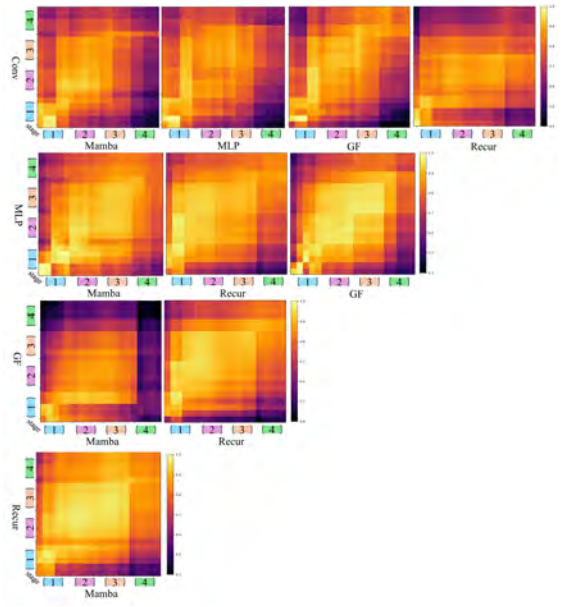

Fig. 4. Pairwise comparison of the learned representational similarity of intermediate features across different token mixers in MetaUNETR utilizing CKA analysis. The x and y axes denote the backbone 4 stages. Similarity scores range from 0.4 to 1, with lighter colors indicating higher similarity. Along the antidiagonal in plots, layer-wise and stage-wise consistencies across backbones are revealed for the top three stages, indicating stage 4 contributes minimally to model performance, rendering it redundant.

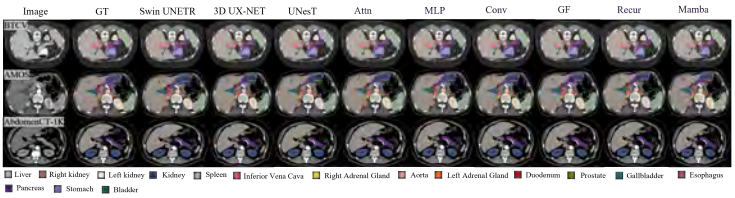

Fig. 5. Qualitative analysis on multi-organ segmentation performance among MetaFormer and prior arts.

Table 2. The overview of key statistics of three public datasets: BTCV, AMOS, and AbdomenCT-1K.

Utilized Dataset	BTCV	AMOS	AbdomenCT-1K
Sample size	30	300	1000
Imaging modality	Portal venous-contrast CT	Multi-contrast CT	Multi-contrast CT
Resolution	512 × 512 × {85 - 198}	{512 - 768} × {512 - 768} × {68 - 353}	512 × 512 × {31 -1026}
Spacing	{0.54 - 0.98} × {0.54 - 0.98} × {2.5 - 5.00}	{0.45 - 1.07} × {0.45 - 1.07} × {1.25 - 5.00}	{0.45 - 1.07} × {0.45 - 3.00} × {0.45 - 8.00}
Anatomical label	spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, right adrenal gland, left adrenal gland	spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus.	spleen, kidney, liver, pancreas

Table 3. The overview of key hyperparameters for data transform and training protocols on three public datasets.

Hyperparameters		BTCV	AMOS	AbdomenCT-1K
Data Transform	Spacing	$1.5 \times 1.5 \times 2mm^3$		
	ScaleIntensity	{Min_value: -170, Max_value: 250}	{Min_value: -125, Max_value: 275}	{Min_percentile: 5th, Max_percentile: 95th}
	RandCropByPosNegLabel	$Patch \quad Size: 96 \times 96 \times 96, background = 3:1$		
	RandShiftIntensity	Offsets:0.10		
Training protocol	Iteration numeber	30000	30000	50000
	Batch Size	1	2	4
	Optimizer	AdamW, beta =(0.9,0.999),eps=$1e-8$, weight_decay=$1e-5$		
	Peak learning rate	$1e-4$		
	Learning Rate Scheduler	Cosing Annealing,T_0 = training epoches,eta_min=$1e-4$		
	Base channel	48		