Supplementary Material for

Revisiting Self-Attention in Medical Transformers via Dependency Sparsification

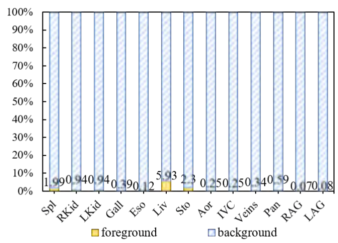

Fig. 1: Foreground ratios of different organs. The foreground region of the 13 organs in BTCV is relatively low compared to the entire image, with the highest proportion being less than 6%.

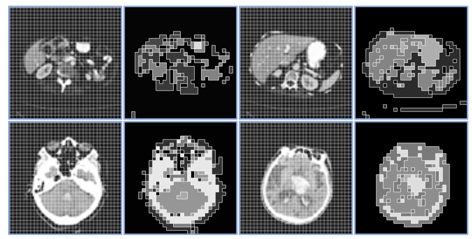

Fig. 2: Dependency merging results on SETR where the original number of tokens is 1024.

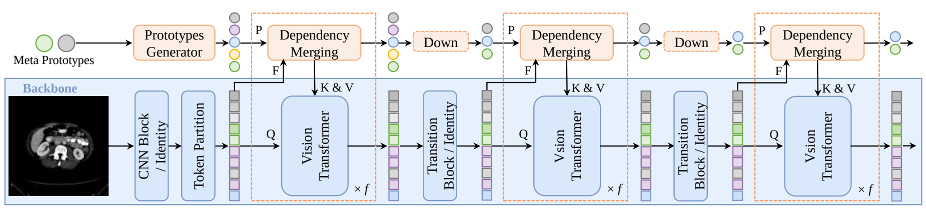

Fig. 3: Example of using DMA as a plug-and-play module in a ViT-based backbone. Given a vanilla ViT, K and V are merged into shorter sequences via dependency merging, and feature prototypes are updated through DMA. After performing f transformer layers, it is optional to decide whether to down-sample the prototypes to further reduce computational complexity and deepen the semantic features of prototypes. The down-sampling operation is realized by averaging the adjacent two prototypes as they inherit from the same parent prototype in the generation process.

Table 1: Ablation study on the factors α and β on TransUNet, evaluated on BTCV. A larger α can better separate the positive and negative prototypes but may turn one of them into the outlier feature embeddings. Given a smaller β, the features within P_+/P_- tend to become similar, causing a decrease in feature diversity. Comparatively, given a larger β, the feature diversity is enriched but may wrongly group different objects into the same prototype.

α	β	Dice	HD	IoU	SE
1	0.05	77.57	19.49	66.54	77.54
1	**0.1**	**78.67**	19.25	**67.93**	**79.67**
1	0.3	77.89	19.19	67.11	77.96
1	0.5	77.56	19.16	66.5	76.95
0.5	0.1	78.43	**18.96**	67.64	77.58
5	0.1	77.61	20.37	65.98	76.95

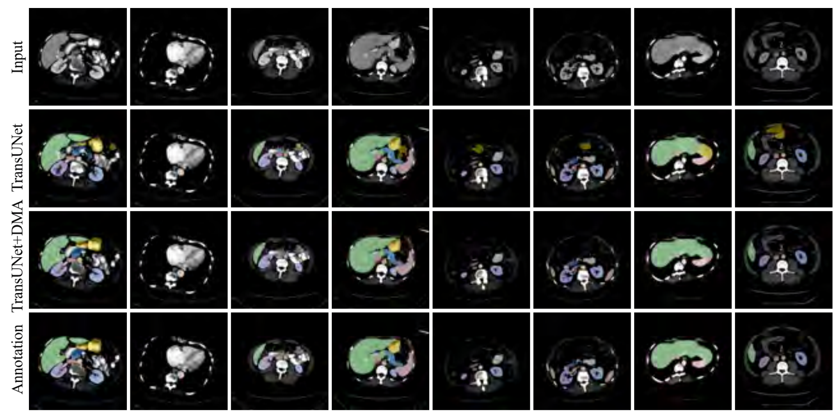

Fig. 4: Qualitative comparisons between TransUNet and TransUNet with DMA on BTCV.

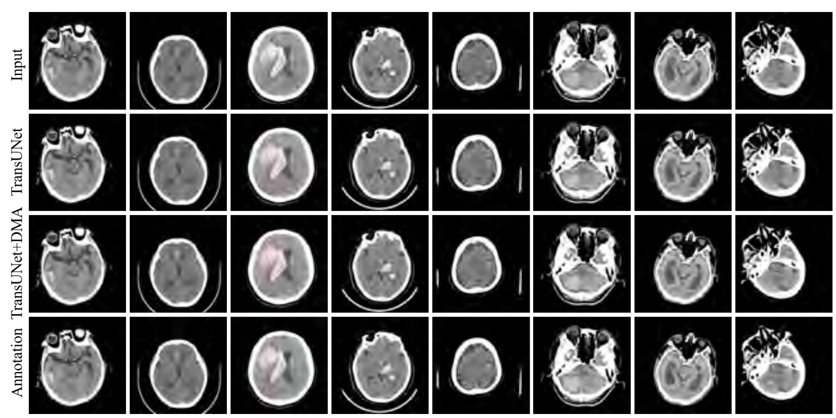

Fig. 5: Qualitative comparisons between TransUNet and TransUNet with DMA on INSTANCE.