Versatile Medical Image Segmentation Learned from Multi-Source Datasets via Model Self-Disambiguation

Supplementary Material

1. More Details about Datasets

Details of the seven public datasets are provided in their corresponding papers. Regarding the private dataset, it comprises 122 contrast-enhanced CT images from patients undergoing urinary system examinations. The images have a uniform matrix size of 512×512, with a variable number of 2D slices ranging from 62 to 685. Pixel spacing ranges from 0.607 to 0.977 mm, and slice thickness varies from 1.0 to 3.0 mm. Urologists annotated the kidney, bladder, and ureters in each image. For this study, only the masks of the two kidneys and the bladder were retained.

2. More Details about Network Architecture

Table 10 presents the architecture for 3D TransUNet. The structure of 3D TransUNet is asymmetric, featuring a greater number of layers in the encoder compared to the decoder. Both the encoder and decoder are composed of 5 stages, wherein spatial sizes progressively decrease by 50% from stage 1 to stage 5 in a sequential manner.

The network's building blocks are shown in brackets. All blocks, except those in stage 5, comprise two consecutive convolutional layers. The adjacent pair of numbers within each bracket represent the input channels and output channels of a convolutional layer. A skip connection [4] is added when the input channels of the first convolutional layer is different from the output channels of the second convolutional layer within each building block in stages 1–4. In accordance with [2], we employed weight normalization [1] in every convolutional layer to expedite training. Subsequent to each convolution operation, instance normalization [3] and rectified linear unit activation are applied. Downsampling and upsampling are executed through trilinear interpolation.

At the bottleneck, four multi-head attention layers were incorporated, each with eight heads. The size of each attention head for query, key, and value was set to be 512.

Table 10. Network architecture.

	Encoder	Decoder
Stage 1	{1, 32, 64}	{128, 64, 64}
Stage 2	{64, 64, 128} {128, 128, 256}	{128, 64, 64}
Stage 3	{256, 128, 256} {256, 128, 256} {256, 128, 512}	{512, 64, 64}
Stage 4	{512, 256, 512} {512, 256, 512} {512, 256, 512} {512, 256, 1024}	{1024, 256, 256}
Stage 5	{1024, 512}	{512, 512}

3. More Results on Partially Labeled Data

Effect of patch size. We conducted experiments with two different patch sizes, namely $96 \times 96 \times 96$ and $112 \times 112 \times 112$. Larger patch sizes were not explored due to limitations in GPU memory. As indicated in Table 11, employing a patch size of $96 \times 96 \times 96$ resulted in an average DSC of 87.9%, which is 0.8% DSC lower than the performance achieved with a patch size of $112 \times 112 \times 112$. These findings underscore the advantageous impact of using a larger patch for abdominal organ segmentation, since increased patch size contributes to a more comprehensive inclusion of contextual information.

Table 11. Performance trained with different patch sizes.

Patch Size	DSC [%]
$96 \times 96 \times 96$	$87.9_{\pm 8.1}$
$112 \times 112 \times 112$	$\mathbf{88.7}_{\pm \mathbf{7.0}}$

Effect of voxel spacing. In our experiments, we standardized the voxel spacing for all images to facilitate model training. To assess the influence of voxel spacing on model performance, we conducted experiments with three different voxel spacings: $1.5 \times 1.5 \times 1.5$ mm^3, $2.0 \times 2.0 \times 2.0$ mm^3, and $2.5 \times 2.5 \times 2.5$ mm^3. As indicated in Table 12, employing a voxel spacing of $1.5 \times 1.5 \times 1.5$ mm^3 and $2.5 \times 2.5 \times 2.5$ mm^3 led to a performance decrease of 0.3% and 0.6% in terms of average DSC, respectively. The diminished performance with a voxel spacing of $1.5 \times 1.5 \times 1.5$ mm^3 can be attributed to reduced contextual information within the input image patch. Conversely, the inferior performance with a voxel spacing of $2.5 \times 2.5 \times 2.5$ mm^3 is likely due to information loss during downsampling, particularly impacting small structure segmentation.

4. More Results on Sparsely Labeled Data

Tables 13 and 14 provide a detailed comparison of the performance for each anatomical structure and across each dataset, respectively. Visual results of a randomly selected

Table 12. Performance trained with different patch sizes.

Voxel Size	DSC [%]
$1.5 \times 1.5 \times 1.5$	$88.4_{\pm 7.7}$
$2.0 \times 2.0 \times 2.0$	$\mathbf{88.7}_{\pm \mathbf{7.0}}$
$2.5 \times 2.5 \times 2.5$	$88.1_{\pm 7.8}$

subject from each dataset are presented in the second and third columns of Fig. 5. These results align with the findings in Table 8, highlighting the consistent success of our method across different views. Notably, even with the utilization of only 20% of incompletely annotated slices for training, our method demonstrates commendable performance across the structures of interest and datasets.

5. More Results on Hybrid Data

Tables 15 and 16 present a comprehensive comparison of performance for each anatomical structure and across each dataset, respectively. Visual results of a randomly selected subject from each dataset are displayed in the fourth column of Fig. 5. These results concur with the findings in Table 9, underscoring the effectiveness of our method in utilizing a mixture of partially and sparsely labeled data for model training.

References

[1] Salimans, T. et al. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems, 29. 1

[2] Chen, Jieneng, et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. 3, 1

[3] Ulyanov, D. et al. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022. 1

[4] He, K. et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). 1

Table 13. Performance (DSC, %) comparison on each anatomical structure using different portions of annotated slices.

Setting	View	Sp	RK	LK	GB	Eso	L	St	A	PC	Pan	RAG	LAG	Duo	B	PU	PSV	Average
20%	Axial	93.5 ±9.3	92.6 ±10.9	92.4 ±9.9	74.3 ±26.4	78.8 ±12.0	95.4 ±9.6	89.5 ±13.7	89.4 ±12.1	83.2 ±15.0	83.1 ±11.4	70.8 ±14.1	72.4 ±16.3	70.5 ±23.5	85.7 ±17.2	71.1 ±22.9	70.7 ±20.1	82.0 ±15.3
	Sagittal	94.5 ±8.1	92.5 ±11.1	91.9 ±9.0	75.6 ±25.5	76.9 ±11.3	96.1 ±7.1	90.5 ±12.2	91.0 ±8.7	85.0 ±12.9	83.5 ±9.3	71.6 ±13.0	73.1 ±14.6	71.7 ±19.7	88.6 ±14.6	74.2 ±24.5	72.3 ±18.5	83.1 ±13.8
	Coronal	94.8 ±6.8	92.6 ±10.8	92.7 ±8.0	75.1 ±24.3	76.5 ±12.6	96.0 ±7.9	90.7 ±12.0	90.6 ±7.8	85.5 ±13.5	83.5 ±10.5	71.3 ±14.6	71.6 ±16.6	73.3 ±18.6	88.2 ±15.3	73.3 ±20.2	74.3 ±17.6	83.1 ±13.6
100%	Axial	95.0 ±7.0	93.5 ±8.4	93.9 ±5.8	76.1 ±26.1	80.7 ±10.2	96.6 ±5.4	91.3 ±11.5	91.7 ±9.0	86.2 ±12.3	84.6 ±8.8	74.5 ±12.0	74.9 ±14.3	75.0 ±18.3	89.2 ±13.6	76.2 ±20.1	76.8 ±15.8	84.8 ±12.4
	Sagittal	94.6 ±8.6	93.1 ±8.7	93.7 ±5.1	76.1 ±25.1	80.4 ±8.6	96.2 ±7.1	91.2 ±12.0	91.1 ±8.8	86.4 ±10.4	84.6 ±9.9	72.7 ±14.2	73.9 ±15.2	73.9 ±18.2	89.0 ±14.4	76.9 ±23.8	75.8 ±17.8	84.3 ±13.0
	Coronal	95.2 ±6.6	93.5 ±7.6	93.7 ±5.6	77.7 ±23.5	81.3 ±8.8	96.5 ±6.4	91.1 ±12.2	91.7 ±9.1	86.4 ±12.0	84.4 ±9.7	73.0 ±13.6	74.6 ±14.7	74.8 ±17.1	89.6 ±12.8	73.7 ±19.1	75.2 ±16.6	84.5 ±12.2

Table 14. Performance (DSC, %) comparison on each dataset using different portions of annotated slices.

Setting	View	AbCT-1K	AMOS-CT	AMOS-MRI	BTCV	FLARE22	NIH-Pan	TotalSeg	Urogram	WORD
20%	Axial	92.7 ±2.0	82.7 ±8.1	80.0 ±10.1	76.3 ±5.8	89.8 ±1.5	83.5 ±4.7	79.3 ±15.2	92.7 ±2.6	82.0 ±4.4
	Sagittal	92.7 ±1.9	83.4 ±7.6	80.5 ±8.5	76.5 ±6.5	89.4 ±1.5	83.2 ±5.3	82.0 ±11.3	92.3 ±3.0	82.7 ±4.4
	Coronal	92.8 ±2.3	83.5 ±7.2	79.2 ±9.8	76.1 ±7.8	89.8 ±1.6	83.3 ±5.8	81.6 ±13.1	92.6 ±3.0	82.5 ±4.2
100%	Axial	93.4 ±1.7	85.1 ±6.3	80.9 ±9.0	77.9 ±6.4	90.5 ±1.8	83.8 ±5.4	84.7 ±9.7	93.0 ±3.0	83.7 ±4.1
	Sagittal	93.0 ±2.2	84.2 ±7.7	81.5 ±7.7	78.7 ±5.5	90.4 ±1.0	84.5 ±5.2	84.2 ±9.8	93.0 ±2.9	83.5 ±4.4
	Coronal	93.3 ±2.0	84.5 ±6.9	81.1 ±8.3	78.1 ±5.7	90.3 ±1.3	84.4 ±4.2	85.0 ±9.5	93.0 ±3.2	83.7 ±4.0

Table 15. Performance (DSC, %) comparison on each anatomical structure using mixed training.

View	Sp	RK	LK	GB	Eso	L	St	A	PC	Pan	RAG	LAG	Duo	B	PU	PSV	Average
Axial	95.1 ±9.3	93.4 ±10.9	93.8 ±9.9	75.8 ±26.4	81.6 ±12.0	96.4 ±9.6	90.7 ±13.7	91.7 ±12.1	86.1 ±15.0	84.5 ±11.4	74.6 ±14.1	75.7 ±16.3	73.4 ±23.5	89.0 ±17.2	76.6 ±22.9	73.2 ±20.1	84.5 ±15.3
Sagittal	95.1 ±7.1	93.2 ±10.3	93.7 ±6.6	77.6 ±24.0	80.6 ±10.7	96.6 ±6.5	91.6 ±11.5	92.1 ±7.1	86.5 ±11.2	84.9 ±8.5	74.1 ±13.7	75.5 ±14.3	75.3 ±17.5	89.2 ±14.7	76.6 ±26.7	75.9 ±17.4	84.9 ±13.0
Coronal	95.1 ±6.8	93.8 ±6.4	93.8 ±6.0	77.2 ±25.0	80.6 ±11.3	96.4 ±7.3	91.3 ±12.4	91.9 ±7.8	86.3 ±12.4	84.7 ±8.9	73.9 ±14.2	75.7 ±13.8	75.1 ±18.2	89.5 ±14.2	77.6 ±23.6	76.7 ±17.4	85.0 ±12.9

Table 16. Performance (DSC, %) comparison on each dataset using mixed training.

View	AbCT-1K	AMOS-CT	AMOS-MRI	BTCV	FLARE22	NIH-Pan	TotalSeg	Urogram	WORD
Axial	93.5 ±1.9	85.1 ±6.5	80.9 ±8.9	76.5 ±5.7	90.8 ±1.0	84.3 ±5.2	84.3 ±10.0	93.3 ±3.0	83.2 ±4.2
Sagittal	93.3 ±1.8	85.1 ±7.4	80.7 ±8.5	77.1 ±7.2	90.8 ±0.9	84.2 ±4.4	84.7 ±10.6	93.1 ±3.1	83.5 ±4.0
Coronal	93.3 ±1.7	85.1 ±7.2	80.8 ±9.3	77.6 ±6.6	90.5 ±1.1	84.5 ±4.8	84.1 ±12.7	93.3 ±2.7	83.6 ±4.1

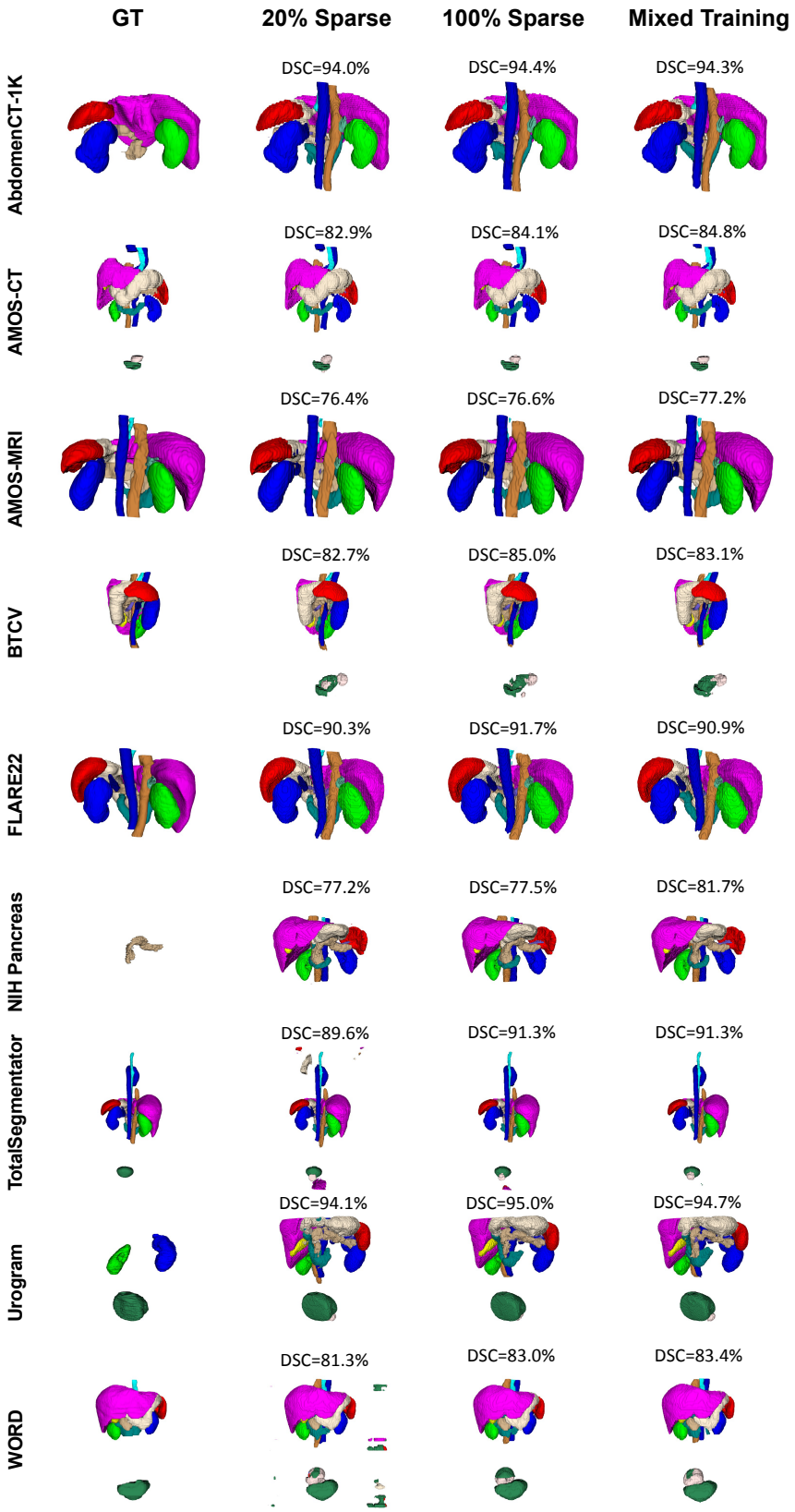

Figure 5. Visual comparisons between the ground truth and predictions from models trained with 20% slices of the axial view, 100% slices of the axial view (loss is computed slice-wise to emulate sparsely labeled data), and hybrid data (the entirety of AMOS, BTCV, and FLARE22 is utilized, while 20% slices of the axial view are taken from other datasets for training) on subjects from various datasets. For a clearer view of detailed differences, zoom in to closely examine the results.