Appendix

Contents

A Abdominal Organ Segmentation Datasets

In this part, we provide detailed descriptions of previous abdominal organ segmentation datasets. We first introduce the datasets covering single organs in Sec. A.1. The introductions of multi-organs Datasets will be developed in Sec. A.2. The statistics of the specific datasets are summarized in Table 8.

A.1 Single-organ Datasets

MSD-Liver dataset [1] consists of 131 training and 70 testing CT cases with liver and liver tumor annotations. These scans were collected at 7 medical centers with patients suffering from various primary cancers.

MSD-Spleen dataset [1] provides 61 cases with spleen annotations, which are provided by the Memorial Sloan Kettering Cancer Center (New York, USA) with patients undergoing chemotherapy treatment for liver metastases. The annotations are first generated using a level-set-based method semi-automatically, and finally revised by an expert abdominal radiologist.

MSD-Prostate dataset [1] maintains 48 prostate multiparametric MRI (mpMRI) with the annotations covering the prostate peripheral zone and the transition zone. The data was acquired at Radboud University Medical Center, Nijmegen Medical Centre, Nijmegen, the Netherlands.

MSD-Pancreas dataset [1] contains 420 patients suffering from pancreatic masses in Memorial Sloan Kettering Cancer Center (New York, USA). The pancreatic parenchyma and pancreatic mass (i.e., cyst or tumor) annotations are provided, which are manually annotated by an radiologist.

KiTS dataset [8] includes 300 cases with kidney and kidney tumor annotations, which are acquired at the University of Minnesota Medical Center (Minnesota, USA). The patients in this dataset are suffered from kidney cancer. The kidney and tumor annotations were segmented by junior medical students under the supervision of a clinical chair.

	Intensivty Property				Spatial Property	
	median	mean	5%	99.50%	slice spacing (mix/max/median)	slice num range (mix/max/median)
MSD-Liver	101	99.39	-17	201	[0.70 / 5.0 / 1.0]	[74 / 987 / 432.0]
MSD-Spleen	105	99.29	-41	176	[1.5 / 8.0 / 5.0]	[31 / 168 / 90.0]
MSD-Prostate	641	854.69	0	2186	[3.0 / 4.0 / 3.6]	[11 / 24 / 20]
MSD-Pancreas	84	77	-96	215	[0.70 / 7.5 / 2.5]	[37 / 751 / 93.0]
Kits	100	100	-79	303	[0.43 / 1.04 / 0.78]	[512 / 796 / 512]
BTCV	96	83	-958	326	[2.5 / 5.0 / 3.0]	[85 / 198 / 127]
Chaos	325	361.38	40	1081	[5.5 / 9.0 / 9.0]	[26 / 50 / 30]
DenseVNet	32	-46	-1003	443	[1.25 / 5.0 / 2.5]	[37 / 174 / 93]
AMOS	-	-	-	-	[0.82 / 6.0 / 5.0]	[40 / 535 / 115.0]
AMOS-CT	57	50	-991	362	[1.25 / 5.0 / 5.0]	[67 / 369 / 115.0]
AMOS-MRI	768	25383	32	164273	[0.82 / 6.0 / 2.0]	[40 / 535 / 115.0]

Table 8: Intensity and Spatial statistics of the conventional abdominal organ segmentation datasets. The slice spacing and number of slices denote the axial plane spacing and resolution of the images

A.2 Multi-organ Datasets

BTCV dataset [12] consists of 50 abdominal CT scans acquired at the Vanderbilt University Medical Center from metastatic liver cancer patients or post-operative ventral hernia patients. This benchmark aims to segment 13 organs, including the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic vein, pancreas, right adrenal gland, and left adrenal gland. The organs were manually labelled by two experienced undergraduate students, and verified by a radiologist.

Chaos dataset [11] consists of 20 CT scans with liver annotations and 20 MRI cases with four organ annotations (i.e., liver, spleen, left kidney, right kidney), which are collected by Dokuz Eylul University (DEU) hospital (˙Izmir, Turkey). The samples in this dataset are acquired from a healthy population.

DenseVNet dataset [5] comprise 90 abdominal CT images and the corresponding segmentation masks of 8 organs. The cases are collected from the Cancer Imaging Archive (TCIA) Pancreas-CT dataset with pancreas segmentation, and the Beyond the Cranial Vault (BTCV) challenge with the segmentation of all organs except the duodenum. An imaging research fellow manually labeled the unsegmented organs under the supervision of a board-certified radiologist.

AbdomentCT-1K dataset [15] is a dataset with 1132 cases covering the liver, kidney, and pancreas annotation, which consists of 1112 3D CT scans from five existing datasets, including MSD-Liver (201 cases), KiTS (300 cases), MSD Spleen (61 cases) and Pancreas (420 cases), NIH Pancreas (80 cases), and a new dataset from Nanjing University (50 cases). Specifically, the overall 50 CT scans in the Nanjing University dataset are from 20 patients with pancreas cancer, 20 with colon cancer, and 10 with liver cancer. Annotations from the existing datasets are used if available. Besides, the absent organs will be further annotated in these datasets.

Word dataset [14] consists of 150 cases with 16 types of organ annotations. The scans are collected from patients who had the prostatic cancer, cervical cancer, or rectal cancer. The annotations were manually labeled from scratch.

Slice spacing and number of slices refer to the 3D image axial plane resolution and spacing

B AMOS: Additional Details

B.1 Data Acquisition

All data in AMOS are collected from eight scanners with different brands. Acquisition details are different for each institution since they follow different clinical protocols in the clinical scenario. For example, 50 CT scans collected from the same scanner are obtained via the criteria of 120 kVP tube, 500 mm data collection diameter, 500-800 ms exposure time, and 50-400 mA Xray tube current. Images were reconstructed at the 2.5-5 mm section thickness with a standard FC08 convolutional kernel and a 400-500 mm reconstruction diameter. All data contributions to this study have been reviewed and approved by the Research Ethics Committee of Longgang District People's Hospital

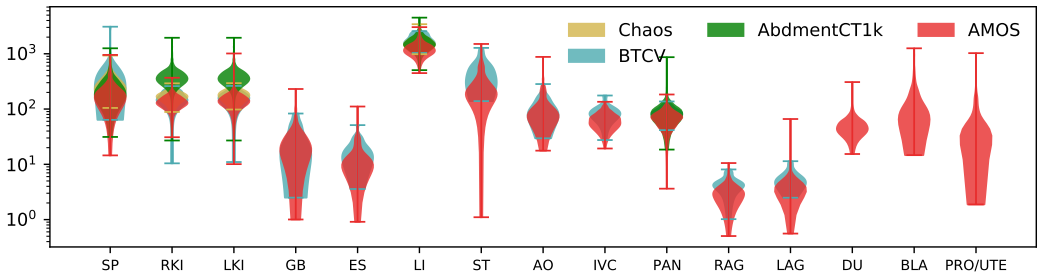

Figure 4: Organ volume distribution of BTCV, Chaos, AbdomentCT-1K and AMOS datasets.

(reference number: 2021077) and the Research Ethics Committee of Longgang District Central Hospital (reference number: 2021ECJ012). The approved documents can be found in https://drive.google.com/drive/folders/1UNHjEgau85rit-DBAKg9kGv6REkiHU6Z?usp=sharing.

B.2 Data Annotation

AMOS adopts a semi-automatic annotation workflow as shown in Figure 2 in the main paper. The coarse masks generated by a pre-trained segmentation model are first refined by five junior radiologists with 5 years of experience in clinical scanning, and further supervised by three board-certified senior radiologists with 10 years of experiences. The segmentation labeling was performed slice-by-slice in the sagittal plane. Besides, the volumetric consistency was enforced by correcting segmentation in the axial and coronal planes in the ITK-SNAP [26] toolbox. Each scan is annotated by one single annotator without multiple annotations for aggregation. For the consistency of the annotation review, specifically, each senior physician will first individually review and record their comments, including a description of the problem and the corresponding image location, and then the comments will be aggregated and discussed to reach a final consensus opinion.

B.3 Data Distribution, Hosting, and Maintenance

All data is distributed under the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) license. Data is hosted on the AWS open data platform and maintained by the authors. Instructions for downloading and using the dataset can be found in the page https://amos22.grand-challenge.org/. Further, we will establish a github repository to solicit possible annotation errors from data users.

	Intensivty Property				Spatial Property	
	median	mean	5%	99.50%	slice spacing (mix/max/median)	slice num range (mix/max/median)
AMOS-CT-A	69	63.76	-967	403	[5.0/ 5.0/ 5.0]	[68/ 140/ 98]
AMOS-CT-B	45	25	-989	158	[5.0/ 5.0/ 5.0]	[68/ 112/ 93]
AMOS-CT-C	63	53.68	-996	401	[1.25 / 5.0 / 2.0]	[76/ 353/ 213]
AMOS-CT-D	61.0	54	-994	337	[1.25 / 5.0 / 2.0]	[78 / 321 / 203]
AMOS-CT-E	59	56	-979	353	[1.25 / 5.0 / 2.0]	[67 / 369 / 224]
AMOS-MRI-F	57422	58761	59	170721	[0.86/ 6.0 / 1.40]	[60 / 535 / 320]
AMOS-MRI-G	2560	344	16	66331	[0.82 / 3.0 / 3.0]	[64 / 512 / 72]
AMOS-MRI-H	845	22937	47	1451508	[0.82 / 2.0 / 0.82]	[100 / 512 / 512]

Table 9: Intensity and spatial statistics of data generated from different scanners. Unseen test data are marked as gray. The slice spacing and number of slices denote the axial plane spacing and resolution of the images

B.4 Data Statistics

Data were collected from patients with abdominal tumors (majority) or other abnormalities. Moreover, among the 500 CT scans collected, the number of males and females are 314 and 186, respectively. For the age distribution, the patients' minimum, maximum, median, and mean ages are 14, 94, 54, and 53.64 years old, respectively. For the 100 MRI scans, the number of males and females are 55 and 45, and the patients' minimum, maximum, median, and mean ages are 22, 85, 50, and 48.71 years old, respectively. The ratio between the number of patients diagnosed with tumors and the

Model	mDice ↑	Categorical Dice ↑														
		SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAG	LAG	DUO	BLA	PRO/UTE
UNet [19]	88.87	96.31	95.29	96.28	81.53	85.72	97.05	90.77	95.37	91.53	87.39	79.83	81.12	82.56	88.42	83.81
VNet [16]	81.96	94.21	91.86	92.65	70.25	79.04	94.65	84.79	92.96	87.4	80.5	72.62	73.19	71.69	77.02	66.62
CoTr [24]	77.13	91.09	87.18	86.36	60.47	80.9	91.61	80.09	93.66	87.72	76.32	73.68	71.74	67.98	67.38	40.84
nnFormer [28]	85.63	95.91	93.51	94.8	78.47	81.09	95.89	89.4	94.16	88.25	85.0	75.04	75.92	78.45	83.91	74.58
UNetr [7]	78.33	92.68	88.46	90.57	66.5	73.31	94.11	78.73	91.37	83.99	74.49	68.15	65.28	62.35	77.44	67.52
Swin-UNetr [6]	86.37	95.49	93.82	94.47	77.34	83.05	95.95	88.94	94.66	89.58	84.91	77.2	78.35	78.59	85.79	77.39

Model	mNSD ↑	Categorical NSD ↑														
		SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAD	LAD	DUO	BLA	PRO/UTE
UNet [19]	79.87	89.62	88.89	89.86	72.91	78.41	83.9	76.29	90.61	79.62	73.37	83.32	83.25	69.35	76.16	62.43
VNet [16]	67.94	83.39	81.14	83.59	58.43	65.56	74.99	63.49	84.86	69.62	60.29	74.23	72.33	52.63	56.8	37.7
CoTr [24]	64.15	77.66	75.27	74.64	48.97	69.47	71.91	59.46	86.34	72.13	56.82	76.18	70.68	52.39	54.52	15.81
nnFormer [28]	74.15	87.73	85.82	87.41	68.0	69.85	80.88	72.3	86.82	71.73	68.21	77.95	77.32	61.09	67.7	49.39
UNetr [7]	61.49	77.58	76.06	77.36	50.78	58.66	72.09	51.45	78.89	60.91	51.91	69.74	63.23	41.16	55.28	37.3
Swin-UNetr [6]	75.32	87.46	85.7	86.76	67.34	73.62	80.81	71.17	88.85	74.99	68.36	80.52	79.65	61.97	69.38	53.2

Table 10: The class-wise scores on the validation set of AMOS-CT dataset.

number of patients with other abnormalities is 3:2. We manually set the distribution of these factors consistent between the training/validate/test splits. We also analyze the intensity and spatial property of the data collected from different scanners and summarize them in Table 9.

C Experiment

In this section, we present the experimental details covering the model architectures, training schedules and so on. We follow the nnUNet package [9] to conduct the model training and evaluation in Pytorch [18]. The code used to produce the results in our paper will be available at https://github.com/JiYuanFeng/AMOS2022.

C.1 Implementation Details

Data Prepossessing Following [9], for the CT data, we first clip the HU values of each scans to the [-991, 362] range and then normalize truncated voxels values by subtracting 50 and dividing by 141. As for the MRI data, we adopt Z-score data normalization.

Baselines We benchmark various state-of-the-art medical segmentation methods. Unless otherwise specified, we follow the default configurations in their released codebases. The implementation of these methods can be found in: UNet[2], VNet[3], CoTr[4], nnFormer[5], UNetr[3], Swin-UNetr[3].

Parameter	Prob	Param
Random Rotation	0.2	[-0.52, 0.52]
Random Scale	0.2	[0.70, 1.40]
Random Gaussian-Noise	0.1	[0.00, 0.10]
Random Gaussian-Blur	0.2	[0.50, 1.00]
Random Brightness	0.15	[0.75, 1.25]
Random Contrast	0.15	[0.75, 1.25]
Simulate Low-Resolution	0.25	[0.50, 1.00]
Random Gamma	0.3	[0.7, 1.5]
Random Mirror	1	

Table 11: Parameters of the used data augmentations

Training Schedule In the training stage, we randomly crop sub-volume sizes to $64 \times 160 \times 160$ and $48 \times 160 \times 224$ for CT, and MRI scans as input, respectively. All the experiments are conducted using 1 NVIDIA V100 GPU with a batch size of 2. For data augmentation, we follow the configurations in [9], including random rotation, scaling, flipping, Gaussian noising, Gaussian blurring, brightness and contrast adjusting, simulation of low resolution, and Gamma transformation. The detailed augmentation parameters are listed in Table 11. We train each model for the same 1000 epochs for fair comparisons. For network optimization, we configured the training objective as the combination of cross-entropy loss and dice loss. Besides, we adopt the SGD algorithm with a momentum of 0.99 and an initial learning rate of 0.01 as the optimizer.

In the testing stage, we employ the sliding window inference strategy where the window sizes equal the training patch size. Besides, data augmentation, like flipping, is also utilized in the testing process.

[2] https://github.com/MIC-DKFZ/nnUNet/blob/master/nnunet/network_architecture

[3] https://github.com/Project-MONAI/MONAI/blob/dev/monai/networks/nets

[4] https://github.com/YtongXie/CoTr/blob/main/CoTr_package/CoTr

[5] https://github.com/282857341/nnFormer/blob/main/nnformer/

Dataset	Modality	Classes	Description	Available	Train	Val	DS Str	Patch
MSD-Liver	CT	2	Liver and tumour	131	104	27	[5, 5, 5]	[128, 128, 128]
MSD-Spleen	CT	1	Spleen	41	32	9	[4, 5, 5]	[64, 192, 160]
MSD-Pancreas	CT	2	Pancreas and tumour	281	224	57	[3, 5, 5]	[40, 224, 224]
MSD-Prostate	MRI	2	Prostate central gland and Peripheral zone	32	25	7	[2, 6, 6]	[20, 320, 256]
MSD-Kits	CT	2	Kidney and tumour	210	168	42	[5, 5, 5]	[128, 128, 128]
MSD-Cardiac	MRI	1	Left Atrium	20	16	4	[4, 5, 5]	[80, 192, 160]
MSD-HepaticVessel	CT	2	Hepatic vessels and tumour	303	242	61	[4, 5, 5]	[64, 192, 192]
ACDC	MRI	4	RV, MLV, LVC	200	160	40	[2, 5, 5]	[20, 256, 224]
Covid-19	CT	1	Covid Lesion	199	159	40	[2, 6, 6]	[28, 256, 256]
SegTHOR	CT	1	Esophagus, Heart, Trachea, Aorta	40	32	8	[4, 5, 5]	[64, 192, 160]

Table 12: Characteristics of the datasets used in transfer learning. Based on the available data, We divide the training and validation set according to the ratio of 8:2. Besides, we report the downsample stride (abbreviated as DS Str) of the used UNet architecture configuration, as well as the input patch size of each task.

.

To quantitatively evaluate the segmentation results, we calculate the Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD) scores. A higher score indicates a better segmentation performance.

C.2 Additional Results

Class-wise results We provide the detailed class-wise scores of the benchmarked methods on the validation set in Table 10. The corresponding abbreviations are presented as follows: spleen (SPL), right kidney (RKI), left kidney (LKI), gallbladder (gbl), esophagus (ESO), liver (LIV), stomach (STO), aorta (AOR), inferior vena cava (IVC), pancreas (PAN), right adrenal gland (RAG), left adrenal gland (LAG), duodenum (DUO), bladder (BLA), prostate/uterus (PRO/UTE).

Transfer learning We perform the task-transfer by fine-tuning the pre-trained models on the ten medical segmentation datasets, including six related datasets containing organ annotations in AMOS, and four unrelated ones. Information about the datasets are summarized in Table 12. We adopt the standard fine-tuning protocols, where we initial the network with the parameters of the pre-trained representation from AMOS. We apply the same training and testing schedule as introduced in Appendix C.1.