

Supplementary Information (SI)

This supplementary information presents the detailed specifications of the datasets accompanying the paper “**A Comprehensive Review of Benchmark Datasets for Deep Learning-based Medical Image Segmentation**”, and all the supplementary datasets are consistent with the classification framework in the main text of the paper for ease of retrieval.

A. Supplementary Datasets

2.1 Organ/Tissue Segmentation Datasets

2.1.1 Head and neck segmentation datasets

(2) Eye segmentation datasets

RIM-ONE for Deep Learning (RIM-ONE DL) is a dedicated dataset for glaucoma assessment co-developed by Spanish research teams. It consists of 485 color fundus images collected from 3 hospitals, including 313 cases of healthy samples and 172 cases of glaucoma data. The annotations include diagnosis labels verified by two experts as well as pixel-level masks of the optic disc and optic cup, supporting both glaucoma binary classification and optic disc-optic cup segmentation tasks. Drishti-GS Retinal Image Dataset for Optic Nerve Head Segmentation (DRISHTI-GS) is an optic nerve head segmentation dataset co-developed by Indian research teams. It contains 101 color fundus images, with the optic disc and optic cup annotated by 4 glaucoma specialists with varying levels of experience using dedicated tools. Complete annotations are only provided for the training set. While the dataset features highly professional annotations and high clinical authenticity, it has limitations such as a small sample size, single imaging modality, and insufficient ethnic representativeness.

2.1.2 Thoracic segmentation datasets

(1) Lung segmentation datasets

The JSRT dataset contains 247 grayscale images from Japanese patients. Some of these images suffer from cardiac overlap, while featuring minimal resolution variation. The annotations are completed by Japanese radiology experts to accurately distinguish the left lung, right lung, and background. The NLM-MC dataset is constructed by Montgomery County, USA. It contains 138 posteroanterior thoracic images, including 80 normal cases and 58 tuberculosis cases. The NLM-SZ dataset is constructed by the Third People’s Hospital of Shenzhen, China. It contains 566 PA thoracic images, with annotations covering a variety of pulmonary diseases. The aforementioned datasets provide a standardized evaluation basis for diagnostic algorithms targeting pulmonary diseases such as pneumonia and pulmonary nodules.

(2) Cardiac segmentation datasets

ETAB is a dedicated benchmark dataset for the field of echocardiography. It is integrated from 5 public ultrasound datasets, covering both healthy and diseased cohorts, and includes multi-view 2D grayscale ultrasound images such as the apical four-chamber view and short-axis view. The annotations are pixel-level structure masks for the left ventricle, myocardial wall, and other structures (with some including ejection fraction and ventricular volume data), which can be used to evaluate the generalization ability of segmentation models across different devices and views. The MSD Heart dataset contains 3D Cine-MRI volumetric data from 30 patients. Each volume includes multiple frames covering the entire cardiac cycle, which can clearly visualize the systolic and diastolic motion of the left atrium. As a thin-walled structure with a wall thickness of 1-3 mm, the left atrium presents high segmentation difficulty in some samples due to blurred boundaries between the left atrium and pulmonary veins, as well as motion artifacts.

(3) Airway segmentation datasets

EXACT’09 is the first benchmark dataset sourced from 8 international institutions, containing 40 cases of CT data. Although it only covers segmental bronchi and larger airway structures and has a small sample size, it has significantly advanced the development of airway segmentation techniques toward precise diagnosis and treatment.

2.1.4 Vascular segmentation datasets

(1) Peripheral Vascular segmentation datasets

The Hand Vascular Near-Infrared Blur (HVNIR) is a dedicated private dataset containing 100 near-infrared blurred images of hand blood vessels. The images are captured using a light source with a wavelength of 750

nm. The diameter of the blood vessels in the images is only 2-5 pixels, and some vascular regions are blurred and disconnected. The annotations are completed by vascular surgeons with reference to anatomical structures, and the boundary error of blurred regions is less than 2 pixels. This dataset is used to verify the segmentation performance of models for hand blood vessels in near-infrared blurred images.

(5) Placental vascular segmentation datasets

FetReg is a dedicated dataset for the FetReg2021 Challenge and also the first large-scale multicenter image dataset of fetoscopic laser photocoagulation. It is derived from 2,718 pixel-level annotated images obtained from 24 fetoscopic surgeries for twin-twin transfusion syndrome. The annotations are categorized into four classes: background, blood vessels, fetus, and surgical instruments. This dataset not only supports the task of placental vessel segmentation, but also serves for the joint task of surgical instrument localization.

2.2 Lesion Region Segmentation Datasets

2.2.1 Head and neck lesion segmentation datasets

(1) Brain lesion segmentation datasets

Brain is a public dataset constructed by Duke University. It contains 110 cases from the cancer imaging archive, with each case having 40-176 slices; after excluding negative samples, 1,168 slices are used for research. The dataset is divided into training, validation, and test sets at a ratio of 7:1:2, and it presents the challenge of complex and irregular tumor boundaries. MSD2019Task-01 is a subset of the Medical Segmentation Decathlon benchmark, which includes 484 samples covering 4 MRI modalities and focuses on testing the generalization ability of models under few-shot settings. The Brain Tumor FLAIR dataset contains FLAIR-sequence MRI images from 200 glioma patients. Annotations are initially labeled by medical students and then reviewed by neuroradiologists. This dataset presents challenges such as blurred tumor boundaries and a natural grayscale transition between peritumoral edema regions and normal cerebral white matter.

The Pituitary Microadenoma Dataset from Ruijin Hospital, Shanghai, and the Pituitary Adenoma Dataset from Quzhou People's Hospital both serve the task of pituitary-related lesion segmentation. The former is a pituitary gland segmentation dataset involving pituitary microadenoma lesions. It is derived from 256 coronal slices of T1-weighted brain MRI scans from 86 patients, and all annotations are completed by medical experts. The latter is a private clinical dataset designed for pituitary adenoma segmentation. Skull plain X - rays are used for small tumors, while CT scans are adopted for large tumors, with the latter having a more complex background.

As a public dataset exclusively designed for the MICCAI challenge, HECKTOR provides fused information of anatomical structures (from CT) and metabolic activity (from PET) via the combination of PET-CT dual-modality images, and is used to verify the accuracy of algorithms in segmenting tumors within complex head and neck anatomy.

(3) Thyroid lesion segmentation datasets

TNSCUI contains 4,554 images, which are divided into a training set of 3,644 images (including 1,641 benign cases and 2,003 malignant cases) and a test set of 910 images at an 8:2 ratio. The images are affected by speckle noise and hyperechoic interference. The DDTI dataset is derived from 298 B-mode ultrasound images of 270 female and 29 male patients. The nodule regions account for only 5%-15% of the images, with a pixel-level imbalance ratio ranging from 1:6 to 1:20.

2.2.2 Thoracic segmentation datasets

(5) Mediastinal lymph nodes segmentation datasets

The Lung Image Segmentation thoracic X-ray is a public dataset jointly derived from the Tuberculosis Control Program in Montgomery County, USA, and the Third People's Hospital of Shenzhen, China. It contains 704 thoracic X-ray images with annotations completed by radiologists, which distinguish lung regions from the background. This dataset places high demands on the global contour capture capability of algorithms, and is particularly suitable for evaluating the ability of algorithms to handle complex scenarios such as blurred lung boundaries and cardiac occlusion.

2.2.4 Breast tissue lesion segmentation datasets

BUSI is a 2D breast ultrasound tumor dataset, derived from 630 images of 600 patients. The images are characterized by high levels of noise and artifacts, making it suitable for verifying algorithm performance in ultra-low data scenarios. BUSB is a 2D breast ultrasound lesion dataset containing 163 images, which is divided into a 75% training set and a 25% test set. Characterized by an extremely low proportion of lesions as well as heavy image

noise and artifacts, this dataset is suitable for verifying the capability of algorithms to segment lesions with low contrast and blurred boundaries.

2.3 Multi-class Segmentation Datasets

2.3.1 Multi-organ/tissue segmentation datasets

(1) Head and neck segmentation datasets

1) Brain segmentation datasets

FeTA2022 is released by the Fetal Brain Tissue Annotation Challenge. It contains super-resolution reconstructed brain MRI images of 280 fetuses at gestational weeks 18-35 from 4 centers, divided into a training set of 120 cases and a test set of 160 cases, with annotations for 7 types of brain tissues and 1 type of background. It is used to evaluate the cross-center generalization ability of multi-category tissue segmentation algorithms for fetal brain. iSeg-2019 is an infant brain MRI segmentation dataset released by the iSeg Challenge under MICCAI. It contains brain data of 39 6-month-old infants from 4 centers, divided into a training set of 10 cases, a validation set of 13 cases, and a test set of 16 cases, with annotations for three types of brain tissues: white matter, gray matter, and cerebrospinal fluid. The dataset is intended to address the problem of cross-center imaging heterogeneity. The cSeg-2022 dataset is derived from the MICCAI Infant Cerebellum Segmentation Challenge. It contains 33 cases of 3T T1-weighted MRI images, divided into a training set of 13 cases and a test set of 20 cases, with annotations for the complete cerebellar structure. It constructs extreme domain shift scenarios using age differences to verify the robustness of algorithms to the dynamic changes in infant brain development.

(2) Thoracic segmentation datasets

3) Cardiac segmentation datasets

The M&Ms dataset contains 375 cases of cardiac MRI data from multiple centers and various vendors. It is used to provide prior shape information of cardiac structures for the ACDC dataset, thereby addressing the problem of structural information deficiency in weakly supervised segmentation. Cardiac is a public dataset dedicated to cardiac MRI segmentation. It contains data from 20 patients; after slicing, 1,293 2D images are actually utilized, with additional 50-shot and 100-shot sub-datasets established. This dataset is designed to verify the performance of cardiac structure segmentation algorithms for short-axis cardiac images. MMWHS is a multi-modal public dataset for whole-heart segmentation. It has annotated 7 categories of cardiac substructures in CT images, including LV, RV, left atrium (LA), right atrial cavity (RA), ascending aorta (AsA), pulmonary artery (PuA), and MYO. This dataset is suitable for semi-supervised complex multi-category segmentation tasks and facilitates cardiac surgical planning.

(3) Abdominal segmentation datasets

1) Abdominal multi-region segmentation datasets

Both CHAOS and CHAOS-T2 are abdominal MRI/CT multi-organ segmentation datasets. The former focuses on the T1-weighted modality, and innovatively introduces cross-modality shape priors from AbdomenCT-1K to address the issue of insufficient MRI supervision signals, supporting weakly supervised learning evaluation. The latter is targeted at the T2SPIR sequence, specifically adapted to few-shot learning scenarios, and its annotations focus on the differentiation of fine organ structures. As a large-scale abdominal CT dataset, AbdomenCT-1K's core value lies in providing shape prior information for the liver, kidneys, and spleen. It constructs shape descriptors via skeleton extraction and log-polar coordinate histograms, offers structural regularization constraints for cross-modality segmentation of MRI datasets such as CHAOS, and addresses the problem of blurred organ boundaries in MRI images. MAALC is a classic abdominal CT multi-organ segmentation dataset derived from an early MICCAI challenge. Although its sample size is only 30 cases, it has become a benchmark for verifying few-shot segmentation performance across CT and MRI modalities due to its consistent few-shot settings with CHAOS-T2. Additionally, its annotations need to address the interference of intestinal gas and adipose tissue on organ boundaries. The WORD dataset is co-constructed by the University of Electronic Science and Technology of China and other institutions. It contains 150 pre-radiotherapy CT scans from West China Hospital, which are divided into 100 cases for training, 20 for validation, and 30 for testing. The dataset has annotated 16 categories of abdominal organs including femoral heads and is applicable to radiotherapy planning scenarios. AMOS is a large-scale abdominal multi-organ dataset supporting both CT and MRI dual modalities. It contains annotations for 15 categories of organs such as the prostate/uterus, with 480 cases of CT and 180 cases of MRI data. It features a dual-task design including Task1 (CT) and Task2 (CT and MRI). The number of liver annotations is 200 times that of adrenal gland annotations, presenting a long-tail distribution, and it is specifically designed to evaluate the generalization ability of cross-modality and multi-organ segmentation. m2caiSeg is the only dataset focusing on laparoscopic surgery scenarios. It extracts 307 images from surgical video frames, covering 13 categories of objects including

organs, surgical instruments, and fluids. It needs to address dynamic object occlusion such as instruments blocking organs and blood blocking instruments, as well as uneven illumination issues, and is specifically designed to evaluate the segmentation robustness of models in dynamic surgical scenarios. The BTCV dataset serves as a benchmark dataset for evaluating interactive algorithms in abdominal CT multi-organ segmentation. It involves annotations for 13 categories of organs across 30 samples, and is applied to the validation of interactive segmentation tools. It is required to support real-time adjustments and the capture of small fine structures such as the gallbladder and pancreas.

(4) Whole-body segmentation datasets

The CTSpine1K dataset focuses on the fine structures of the spine and contains 1,005 cases of data. It classifies 25 categories including vertebral bodies and their accessories ranging from the cervical vertebrae to the sacral vertebrae. The images have no obvious anisotropy, with blurred boundaries between vertebral segments and small-sized accessories. This dataset is applied to evaluate the segmentation accuracy of the skeletal system and supports the diagnosis of spinal diseases as well as spinal surgery navigation. The CT-ORG dataset focuses on 6 categories of key organs such as the liver and lungs, and contains 140 cases of multi-center and multi-scanning condition data. It adopts a hybrid annotation method combining automatic and manual labeling. The dataset features significant differences in organ sizes and obvious class imbalance, serving as a standardized benchmark for multi-organ joint segmentation models and facilitating the diagnosis of multi-organ diseases.

2.3.2 Multi-lesion region segmentation datasets

(1) Head and neck multi-lesion region segmentation datasets

1) Brain multi-lesion region segmentation datasets

BraTS2019 contains 335 cases of multi-modal MRI volume data (each case includes 4 modalities), with annotations for three types of regions: tumor core, edema region, and enhancing tumor. BraTS2021 contains 1251 cases of multi-modal MRI volume data, with additional rare tumor subtypes included. The images are 3D multi-modal MRI volumes, covering four modalities: T1, T1ce, T2, and FLAIR, and are annotated with pixel-level three-class masks for tumor core, edema region, enhancing tumor/background.

2) Intracranial multi-lesion region segmentation datasets

The INSTANCE2022 dataset is the first public 3D anisotropic ICH segmentation dataset released by the INSTANCE challenge under MICCAI. It contains data from 200 patients treated at Peking University Shougang Hospital between 2017 and 2019, covering 5 subtypes of ICH. The annotations are completed by 3 senior physicians following procedures such as rough annotation, fine revision, and consensus confirmation, and the dataset is open-sourced via the Grand Challenge platform. It aims to address the issues of traditional algorithms ignoring anisotropy and the lack of unified evaluation benchmarks, and supports the research and development of anisotropic data processing technologies. The ICH dataset is constructed by teams from Harbin Institute of Technology, Tsinghua University, and other institutions. It contains 250 cases of emergency data from domestic top-tier tertiary hospitals, also covering 5 subtypes of ICH with some cases showing signs of hematoma expansion. The images have no obvious metal artifacts and have undergone preprocessing steps such as skull stripping and window adjustment dedicated to intracranial hemorrhage. Annotations are completed by 3 physicians using ITK-SNAP. This dataset supports the development of hematoma expansion prior segmentation algorithms and addresses the accuracy and uncertainty issues caused by the variable morphology and blurred boundaries of ICH.

(2) Thoracic multi-lesion region segmentation datasets

The RSPECT dataset is jointly constructed by the Radiological Society of North America in collaboration with institutions from 5 countries worldwide. It contains 12,195 cases of multicenter CTPA data, covering central/peripheral embolism, acute and chronic embolism, and right ventricular strain. Its data are annotated with 14 categories of clinical labels by 86 thoracic radiology experts. Although it lacks pixel-level segmentation and follow-up data, it remains a core benchmark for evaluating the generalization of multicenter algorithms and clinical risk stratification. The CTPA dataset is constructed by Uppsala University in Sweden and other institutions. It contains 1,504 cases of non-electrocardiogram-gated CTPA data, divided into an internal set of 700 cases and an external set of 804 cases. It covers subsegmental microembolism, with 18% of cases affected by contrast agent inhomogeneity. The two datasets provide support for the research and development of automated diagnosis algorithms for pulmonary embolism from the dimensions of multicenter screening and clinical-grade precise segmentation, respectively, and promote the intelligent development of diagnosis and treatment.

(4) Skin multi-lesion region segmentation datasets

The PH2 dataset contains 200 images of melanocytic lesions, which are classified into benign nevi, atypical nevi, and melanoma based on lesion types. It is often used in the mode of pre-training on the ISIC2016 training set and full-scale testing on the PH2 dataset to verify the generalization ability of models. The DermIS dataset

covers images of various dermatological diseases, includes detailed clinical metadata and pixel-level annotations, and supports research on segmentation and diagnosis of multiple lesion types. The HAM10000 dataset contains 10,000 multi-source dermoscopic images covering 7 types of skin lesions. It is used to verify the generalization ability of models in scenarios without interactive prompts, and helps address the problems of domain differences between natural and medical images as well as blurred lesion boundaries.

B. Supplementary Tables

This paper summarizes the core characteristics of some organ/tissue, lesion region, multi-organ/tissue, multi-lesion region, multi-organ/tissue and lesion regions segmentation datasets in terms of annotation quality, data distribution, and domain characteristics. It conducts an analysis across dimensions including annotation reliability, inter-observer variability, domain shift risk, class imbalance, and annotation noise, providing crucial reference for subsequent model selection, data preprocessing, and algorithm optimization. The Specific details are provided in Table 1-5.

Table 1: Detailed information of medical organ/tissue segmentation datasets

Dataset	Annotation Reliability	Reliability	Inter-Observer Variability	Domain Shift	Class Imbalance	Labeling Noise
REFUGE	$\kappa=0.83$	(Expert Annotation + Quality Control)	Low ($\kappa=0.83$)	Low	Moderate (Optic Cup/Disc Region: 6%-8%)	≤ 1 pixel, no obvious noise
REFUGE2	Dice=0.88		Low ($\kappa=0.86$)	Low	Moderate (Disc Region: 8%-10%)	≤ 1 pixel, no obvious error
STS-Tooth	$\kappa=0.85$		Low ($\kappa=0.85$)	Low	Moderate (Tooth Region: 12%-16%)	≤ 1 pixel, high annotation consistency
PARES-22	Dice=0.87		Low ($\kappa=0.84$)	Medium	Mild (Multi-organ Aggregate: 25%-30%)	≤ 1 voxel, no confusion in multi-structures
EchoNet-Dynamic	Dice=0.85		Low ($\kappa=0.83$)	Medium	Mild (Myocardium Region: 15%-20%)	≤ 2 pixels, slight motion artifact interference
ETAB	$\kappa=0.80$		Medium ($\kappa=0.78$)	Medium	Severe (Tumor Region: $<4\%$)	≤ 1 voxel, slight edge mislabeling
AIIB2023	$\kappa=0.937$		Low	Medium	Severe (Airway Region: $<1.2\%$)	≤ 1 voxel, mislabeling ratio $<1\%$
ATM22	Dice=0.92 (Preliminary Prediction + Doctor Optimization)		Low ($\kappa=0.88$)	Medium-High	Moderate (Airway Region: 5%-8%)	≤ 1 voxel, slight fracture in distal branches
EXACT'09	TD=85%, BD=80%		Medium ($\kappa=0.76$)	Medium	Severe (Airway Region: $<3\%$)	≤ 3 pixels, edge blur mislabeling
AVT	Dice=0.91 (Expert Annotation)		Low ($\kappa=0.86$)	Medium	Severe (Vessel Region: 1.8%)	≤ 1 voxel, no false positives
FetReg	Dice=0.79		Medium ($\kappa=0.75$)	High	Severe (Instrument Class: $<3\%$, Fetus Class: $<2\%$)	≤ 3 pixels, slight adhesion mislabeling

Table 2: Detailed information of medical lesion region segmentation datasets

Dataset	Annotation Reliability	Inter-Observer Variability	Domain Shift	Class Imbalance	Labeling Noise
BT-Seg	Dice=0.845 (Manual Annotation + Pathological Verification)	Low ($\kappa = 0.82$)	Low	Severe (Tumor Core: 1.8% + Edema: 3.2%)	≤ 1 voxel, extremely low noise after preprocessing
TN3K	Dice=0.83 (Physician Review)	Medium ($\kappa = 0.76$)	Medium-High	Severe (Thyroid Nodule Ratio: 3.5%)	≤ 2 pixels, slight mis-labeling in multi-nodule adhesion regions
TG3K	Dice=0.85 (Standardized Annotation for Glandular Structures)	Low ($\kappa = 0.83$)	Low	Moderate (Thyroid Gland Ratio: 15%)	≤ 1 voxel, no obvious noise
DDTI	Dice=0.79 (Physician Annotation)	Medium ($\kappa = 0.75$)	Low	Severe (Thyroid Nodule Ratio: 2.8%)	≤ 2 pixels, slight noise in low-contrast regions
FUMPE	Dice=0.86 (Physician Review + Semi-automated Assistance)	Medium ($\kappa = 0.77$)	Medium-High	Severe (PE Region Ratio: 0.8%)	≤ 1 voxel, slight boundary blur in peripheral PE regions
LNQ2023	Dice=0.71 (Multi-expert Verification + Full-node Annotation)	Medium ($\kappa = 0.76$)	Medium-High	Severe (Lymph Node Ratio: 0.6%-1.2%)	≤ 1 voxel, slight boundary blur in fused lymph node regions
CVC-ClinicDB	Dice=0.95 (Physician Annotation)	Low ($\kappa = 0.83$)	Medium	Severe (Polyp Region: 3.0%-3.5%)	≤ 1 pixel, low noise level
CVC-ColonDB	Dice=0.93 (Physician Annotation)	Medium ($\kappa = 0.76$)	High	Severe (Polyp Region: 1.5%-2.0%)	≤ 2 pixels, slight missing leakage noise in small polyp regions
ETIS	Dice=0.94 (Multi-center Physician Annotation)	Medium ($\kappa = 0.75$)	High	Severe (Polyp Region: 2.0%-2.5%)	≤ 2 pixels, noise ratio $< 3\%$ in blurred boundary regions
Kvasir	Dice=0.95 (Manual Physician Annotation)	Low ($\kappa = 0.84$)	High	Severe (Polyp Region: 2.5%-3.0%)	≤ 2 pixels, slight boundary blur in color-fused regions
ABUS2023	Dice=0.61 (Physician Annotation + Cross-validation)	Medium ($\kappa = 0.75$)	Medium	Severe (Tumor Region: 0.5%-1.0%)	≤ 2 pixels, slight artifacts in low-SNR regions
BUSB	Dice=0.91 (Manual Physician Annotation)	Medium ($\kappa = 0.76$)	High	Severe (Tumor Region: 1.2%-1.8%)	≤ 2 pixels, speckle noise ratio $< 5\%$

Table 3: Detailed information of medical multi-organ/tissue segmentation datasets

Dataset	Annotation Reliability	Inter-Observer Variability	Domain Shift	Class Imbalance	Labeling Noise
HaN-Seg	Dice=84.7%-94.4% (Physician Annotation + Manual Correction)	Medium ($\kappa = 0.77$)	High	Severe (Small/OARs < 1%); Mild (Multi-organ Aggregate 10%-15%)	≤ 1.7 mm, metal artifact correction ratio < 2%
FeTA2022	Dice=0.81-0.82 (Global Average), Top-algorithm WM=0.89, Ventricles=0.87	Medium ($\kappa = 0.77$)	High	Severe (GM/Deep GM 1%-3%); Moderate (Ventricle/CSF 5%-8%)	≤ 2.5 mm, slight topological pore noise in GM
iSeg-2019	Dice=0.82-0.86 (WM=0.86, GM=0.82, CSF=0.83)	Medium ($\kappa = 0.76$)	High	Mild (WM/GM 30%-35%, CSF 25%-30%)	≤ 2 mm, slight mislabeling in cortical regions
cSeg-2022(24M)	Dice=0.93 (Physician Annotation)	Low ($\kappa = 0.83$)	High	Moderate (Cerebellar Region 5%-8%)	≤ 2 mm, no obvious noise
WMH	Dice=80% (Test Set Average), Cross-scanner; Dice $\geq 72\%$ (Physician Annotation)	Medium ($\kappa = 0.74$ -0.77)	High	Severe (WMH Lesion 0.2%-0.5%)	≤ 6.3 mm (95%HD), missing lesion ratio < 3%
Teeth3DS	Segmentation Accuracy=81.26%-98.59% (Expert Annotation + Clinical Validation)	Low ($\kappa = 0.85$)	Medium	Severe (Single Tooth 0.5%-1.2%)	$\leq 90\mu m$, no false positive annotations
SegTHOR	Dice=0.82-0.93 (Physician Annotation)	Medium ($\kappa = 0.77$ -0.84)	Low	Severe (Esophagus 2.6%, Trachea 3.7%); Moderate (Heart 70.7%, Aorta 23%)	≤ 0.7 mm, slight edge blur noise in esophagus
CMRxCMotion	Dice ≥ 0.85 (Technician Annotation + Physician Review)	Low ($\kappa = 0.83$)	Low	Moderate (Myocardium + Heart Chamber 5%-8%)	≤ 1 mm, severe image motion artifact ratio < 5%
FLARE21/22	Dice=88.3%-90.9% (Physician Annotation + Final Review)	Medium ($\kappa = 0.75$ -0.86)	High	Severe (Adrenal Gland 0.01%-0.03%); Moderate (Liver 15%-20%)	≤ 1 mm, tiny organ mislabeling ratio < 2%
WORD	Dice=87.41% (Physician Annotation from Scratch + Expert Review)	Medium ($\kappa = 0.73$ -0.88)	Medium-High	Severe (Adrenal Gland 0.05%-0.1%); Moderate (Liver 15%-20%)	≤ 3 mm, tiny organ mislabeling ratio < 2%
BTCV	Dice=85.3% (Standard Competition), 89.9% (Free Competition) (Physician-Guided Annotation)	Medium ($\kappa = 0.72$ -0.88)	Medium-High	Severe (Adrenal Gland 0.1%-0.3%); Moderate (Liver 18%-22%)	≤ 2 mm, slight edge blur in small organs
TotalSegmentator CT	Dice=0.943 (Physician Supervision + Iterative Learning)	Low-Medium ($\kappa = 0.75$ -0.88)	High	Severe (Small Structures < 0.5%); Moderate (Liver 15%-20%)	≤ 1.5 mm, slight missing leakage in colon of 35% cases
TotalSegmentator MRI	Dice=0.839 (Physician + Expert Review)	Low-Medium ($\kappa = 0.74$ -0.86)	High	Severe (Adrenal Gland < 0.3%); Moderate (Liver 12%-18%)	≤ 1.5 mm, slight missing leakage in iliac artery of 11% cases

Table 4: Detailed information of medical multi-lesion region segmentation datasets

Dataset	Annotation Reliability	Inter-Observer Variability	Domain Shift	Class Imbalance	Labeling Noise
L2R-OASIS	PSNR=26.50dB, SSIM=0.9229 (Expert Annotation)	Medium ($\kappa = 0.74$ -0.85)	Medium-High	Severe (Micro/Brain Subregions < 1%); Moderate (Whole Brain 15%-20%)	≤ 1.63 mm, no obvious noise
SD-OCT	Dice=0.78, Retinal Layer=0.73-0.89, Hydrops=0.53 (Expert Annotation + KR+GTDP)	Medium ($\kappa = 0.75$ -0.85)	Medium	Severe (Hydrops < 1%); Moderate (Retinal Layer 15%-20%)	$\leq 3.87\mu m$, slight false positives in hydrops regions
ICH	Dice=0.863 (Physician Annotation + Cross-validation)	Medium ($\kappa = 0.74$ -0.80)	Medium	Severe (Hemorrhage Region < 3%)	≤ 0.42 mm, slight blur in irregular hemorrhage boundaries
INSTANCE	Dice=0.6925 (Physician Annotation + Majority Voting)	Medium ($\kappa = 0.73$ -0.79)	Medium	Severe (Hemorrhage Region < 3%, Micro-hemorrhage < 1%)	≤ 0.42 mm, slight boundary blur in SAH subtypes
ISLES22	Dice=0.75-0.82 (Physician Annotation + Algorithm Pre-segmentation)	Medium ($\kappa = 0.72$ -0.78)	High	Severe (Micro-lesion < 1%; Max Volume: 482.15ml)	≤ 0.88 mm, slight artifact noise in intracranial infarction
ATLASv2.0	Dice=0.67 (Dual-rater Verification + ITK-SNAP Refinement)	Medium ($\kappa = 0.71$ -0.77)	High	Severe (Single-lesion Count < 5%; Min: 0.01%)	≤ 1 mm, slight boundary blur in chronic lesions
RSPECT	Sensitivity=98%, Specificity=90% (nnU-Net + Post-processing, Expert Annotation + Consensus)	Medium ($\kappa = 0.73$ -0.79)	High	Severe (Micro-PE < 0.1%; Max Volume: 36510mm ³)	≤ 0.59 mm, slight false positives in hypo-contrast regions
LNDb	Jaccard=0.48-0.55, Volume Classification; Ac=0.80 (Physician Single Annotation)	Medium ($\kappa = 0.72$ -0.77)	Medium	Severe (Nodule < 1%, Higher ratio for < 3mm)	≤ 0.63 mm, slight boundary blur in small nodules
PI-CAI	AUROC=0.91, Gleason ≥ 2 Cancer Detection Rate 68% (6.8% Higher than Physicians; Gold Standard: Biopsy/Follow-up)	Medium ($\kappa = 0.70$ -0.75)	High	Severe (Clinically Significant Cancer < 10%)	≤ 1 mm, slight false positives in low-signal regions
ULS	Average Dice=0.703 (Annotator + Physician Review)	Medium ($\kappa = 0.70$ -0.76)	High	Severe (Rare Disease Lesions < 0.5%)	≤ 9.58 mm (HD95), slight boundary blur in pancreatic/colonic lesions

Table 5: Detailed information of medical multi-organ/tissue and lesion region segmentation datasets

Dataset	Annotation Reliability	Inter-Observer Variability	Domain Shift	Class Imbalance	Labeling Noise
SegRap2023	OARsDice=76.68%-86.70%, GTVsDice=70.42%-73.44% (Physician Annotation + Review)	Medium ($\kappa = 0.70-0.78$)	Medium	Severe (Small OARs < 1%; GTVs < 5%)	$\leq 3\text{mm}$, slight boundary blur in small OARs
CrossMoDA2023	VSDice=87.3%, CochleaDice=84.1% (Physician Review + Consensus Ruling)	Medium ($\kappa = 0.70-0.78$)	High	Severe (Cochlea < 1%; Tumor Sub-region < 5%)	$\leq 0.43\text{mm}$, slight artifacts in posterior fossa regions
PleThora	Median Dice for Pleural Effusion=0.99, Dice=0.81-0.93 (Physician Review + Automatic Segmentation Correction)	Medium ($\kappa = 0.81-0.99$)	High	Severe (Pleural Effusion 0.5%-3%)	$\leq 5.31\text{mm}$ (95HD), trace effusion with minor mislabeling
LiTS	LiverDice=0.963, TumorDice=0.739 (Physician Annotation + Blinded Review)	Medium ($\kappa = 0.70-0.78$)	High	Severe (Small Tumor < 1%; Liver 15%-25%)	$\leq 1.189\text{mm}$ (ASD), slight missing labeling in small tumors
ATLAS	LiverRMS< 4.5%, TumorRMS< 6.2% (Manual Physician Annotation + Dual-rater Verification)	Medium ($\kappa = 0.72-0.77$)	Medium-High	Severe (Tumor Burden 0.1%-57.0%, 850-fold Volume Difference)	$\leq 2\text{mm}$, slight boundary blur in low-contrast tumors
KiTS19	KidneyDice=0.974, TumorDice=0.851 (Student Annotation + Expert Review + Dual-rater Validation)	Medium ($\kappa = 0.75-0.82$)	Medium-High	Severe (Small Tumor < 1%; Kidney 10%-15%)	$\leq 0.8\text{mm}$, slight missing labeling in small tumors
KiTS21	KidneyDice=0.9748, TumorDice=0.8813, CystDice=0.8710 (Three-stage Annotation + Multi-annotation Fusion)	Medium ($\kappa = 0.76-0.83$)	High	Severe (Cyst < 0.5%, Micro-tumor < 1%; Kidney 10%-18%)	$\leq 0.78\text{mm}$, minor false positives in cyst regions
3D-IRCAB	Expert Semi-automatic Segmentation, Corrected Couinaud Anatomy Segmentation Errors (DICOM+VTK Format Annotation)	Medium ($\kappa = 0.76-0.82$)	Medium	Severe (Tumor < 5%; Liver Main Body 15%-20%)	$\leq 2\text{mm}$, no obvious systematic noise
MSD	Dice=0.85-0.90 (Multi-organ/Lesion)	High ($\kappa = 0.80-0.88$)	High	Severe (Lesion Region 0.8%-3.5%)	$\leq 2\text{voxel}$, slight noise in metal artifacts