Benchmarking of Deep Learning Methods for Generic MRI Multi-Organ Abdominal Segmentation

Deepa Krishnaswamy, Cosmin Ciausu, Steve Pieper, Ron Kikinis, Benjamin Billot, Andrey Fedorov

***Abstract*— Recent advances in deep learning have led to robust automated tools for segmentation of abdominal computed tomography (CT). Meanwhile, segmentation of magnetic resonance imaging (MRI) is substantially more challenging due to the inherent signal variability and the increased effort required for annotating training datasets. Hence, existing approaches are trained on limited sets of MRI sequences, which might limit their generalizability. To characterize the landscape of MRI abdominal segmentation tools, we present here a comprehensive benchmarking of the three state-of-the-art and open-source models: *MRSegmentator*, *MRISegmentator-Abdomen*, and *TotalSegmentator MRI*. Since these models are trained using labor-intensive manual annotation cycles, we also introduce and evaluate *ABDSynth*, a SynthSeg-based model purely trained on widely available CT segmentations (no real images). More generally, we assess accuracy and generalizability by leveraging three public datasets (not seen by any of the evaluated methods during their training), which span all major manufacturers, five MRI sequences, as well as a variety of subject conditions, voxel resolutions, and fields-of-view. Our results reveal that *MRSegmentator* achieves the best performance and is most generalizable. In contrast, *ABDSynth* yields slightly less accurate results, but its relaxed requirements in training data make it an alternative when the annotation budget is limited. The evaluation code and datasets are given for future benchmarking at `https://github.com/deepakri201/AbdoBench`, along with inference code and weights for *ABDSynth*.**

***Index Terms*— benchmark, segmentation, abdominal MRI**

I. Introduction

ACCURATE segmentation of abdominal organs in magnetic resonance imaging (MRI) volumes is a prerequisite for an array of clinical tasks [8] such as volumetry [9], [10], early diagnosis, longitudinal disease monitoring, radiotherapy planning [11], [12], and biomarker extraction [13], [14]. However, manual expert contouring is labor-intensive [15] and prone to inter- and intra-rater reproducibility issues [16]. To mitigate these challenges and improve consistency, automated multi-organ segmentation methods have been proposed.

While automated segmentation of abdominal MRI scans is essential, this task has been hindered by the large variability in acquisition parameters for this modality. Indeed, MRI lacks inherent intensity normalization, which poses a challenge in automatic segmentation since traditional deep neural networks [17] are fragile against MRI contrast variations [18], an issue known as “domain gap” [19]. This variability also complicates manual MRI segmentation and thus hinders the creation of large annotated training sets. As a result, only a few methods have been proposed until recently for multi-organ abdominal MRI segmentation [20]–[23]. This is in contrast with computed tomography (CT), where methods like TotalSegmentator can benefit from the highly standardized signal [3].

Recent advances in deep learning segmentation networks, best represented by nnU-Net [24], have led to the development of state-of-the-art methods for multi-organ abdominal segmentation in MRI: *MRSegmentator* [1], *MRISegmentator-Abdomen* [5], and *TotalSegmentator MRI* [6], each capable of segmenting more than 40 regions, including organs, bones, muscles, and vessels. The development of these methods has been performed jointly with the annotation of their respective training datasets. Specifically, the employed training procedures all rely on a very labor-intensive strategy where annotated datasets are obtained with iterative expert refinements. With this technique, the aforementioned methods are trained on large MRI cohorts from the UK Biobank [2], the German National Cohort (NAKO) [25], Imaging Data Commons [7], TotalSegmentator [3], and other sources [1], [26], [27]. Crucially, these datasets span multiple MRI sequences, manufacturers, voxel resolutions, and pathologies, thereby improving the robustness of these methods compared to previous approaches [20], [21] by increasing the variability of the training data. However, the accuracy and generalizability of these models remain to be compared for generic out-of-the-box usage.

Since direct annotation of MRI datasets is challenging, other approaches propose to tackle the generalization issue of abdominal MRI segmentation networks by adopting domain adaptation strategies [28], where the goal is to transfer models trained on labeled data from a source domain to a target

This work was supported by the NIH NCI (HHSN26110071, HHSN261201500003I) and used Jetstream2 GPU at Indiana University, which is supported by National Science Foundation grants (2138259, 2138286, 2138307, 2137603, 2138296).

D. Krishnaswamy and C. Ciausu contributed equally to this work.

D. Krishnaswamy (dkrishnaswamy@bwh.harvard.edu, corresponding author) is with Brigham and Women’s Hospital, Boston, USA.

S. Pieper (pieper@isomics.com)is with Isomics, Cambridge, USA.

C. Ciausu (cciausu@bwh.harvard.edu), R. Kikinis (kikinis@bwh.harvard.edu), and A. Fedorov (afedorov@bwh.harvard.edu) are with Brigham and Women’s Hospital, Boston, USA.

B. Billot is with Inria, Epione team, Sophia-Antipolis, France (benjamin.billot@inria.fr).

TABLE I
SUMMARY OF THE TRAINING DATA FOR THE BENCHMARKED METHODS. BRACKETS DENOTE RANGES. CE=CONTRAST-ENHANCED.

Method	Training source	# scans	Abnormalities	Data type	Resolution (mm)	Dimension
MRSegmentator [1]	UK Biobank [2]	1200	various pathologies	T1 Dixon in-phase T1 Dixon out-of-phase T1 Dixon water only T1 Dixon fat only	[2.23×2.23×3.00] [2.23×2.23×4.50]	[224×156×44] [224×174×72]
	In-house dataset	221	kidney tumors	T1 T1 fat-saturated T2 fat-saturated	1.00×1.00×1.00	100–450 (only given in axial direction)
	TotalSegmentator [3], [4]	1228	various pathologies	CT	1.50×1.50×1.50	[47×48×29] [499×430×851]
MRISegmentator-Abdomen [5]	In-house dataset	780	liver tumors, pancreatic cysts other abnormalities	T1 pre-contrast CE T1 arterial phase CE venous phase CE delayed phase	[0.94×0.94] [1.47×1.47] (inter-slice res. not given)	[228×240×80] [320×320×96]
TotalSegmentator MRI [6]	University Hospital Basel	1088	various pathologies	T1 T2 Proton density	[0.17×0.17×0.17] [20.0×10.64×14.40]	[11×10×10] [1092×1280×1915]
	Imaging Data Commons [7]		cancer	Various sequences	[0.29×0.29×0.43] [7.50×25.00×28.0]	[17×5×5] [672×672×512]
	TotalSegmentator [3], [4]		various pathologies	CT	1.50x1.50×1.50	[47×48×29] [499×430×851]
ABDSynth	Subset of TotalSegmentator	128	various pathologies	CT segm. (no real images)	1.50×1.50×1.50	300×300×250

domain where labels are unavailable. Li et al. propose training an image-translation network to create synthetic MRI data with pseudo-labels [29]. Segmentation is then performed with an automatic multi-stage network. Another approach [30] first performs CT-to-MRI translation using a CycleGAN model [31] equipped with an organ-attention mechanism. An nnU-Net with some additional enhancements is then trained for supervised segmentation of these MRI scans. However, these domain adaptation approaches need to be retrained for each new MRI sequence, which does not comply with our scenario of out-of-the-box MRI segmentation.

SynthSeg [32], [33] is a relatively new technique that proposes to circumvent domain adaptation strategies with domain randomization [34]. Specifically, SynthSeg leverages a parametric generative model based on a Gaussian mixture model (GMM) conditioned on input segmentations (no real images needed). Synthetic scans are created by randomly sampling all generation parameters from wide uniform distributions, thus yielding scans of randomized MRI contrast. Exposing a downstream segmentation network to such variable scans forces it to learn domain-agnostic features, so that the trained network can be used on any domain without retraining [32]. SynthSeg has originally been proposed for brain segmentation [18], but has since been extended to MRI and CT cardiac segmentation [32]. More generally, the SynthSeg framework is an alternative to the latest methods in MRI abdominal segmentation, since it eases their burdensome annotation process by only requiring segmentations as training inputs, which can be taken from other modalities such as widely available CT label maps [3].

In this paper, we propose a benchmark for existing and future abdominal MRI segmentation methods. In particular, we conduct an analysis on three datasets: AMOS MRI [35], CHAOS MRI [36]–[38], and LiverHCCSeg [39], [40]. These contain a variety of MRI scans acquired at different institutions, using scanners from all major manufacturers, and include variable MRI sequences, voxel resolutions, and populations (i.e., healthy subjects and diseased patients). Here, we assess the performances of the three state-of-the-art methods: *MRSegmentator* [1], *MRISegmentator-Abdomen* [5], and *TotalSegmentator MRI* [6]. For completeness, we also evaluate a new method, named *ABDSynth*, that extends SynthSeg [32] for out-of-the-box MRI multi-organ abdominal segmentation and that is trained solely on CT segmentations. Our results reveal that *MRSegmentator* outperforms the other methods, both for in-domain accuracy and out-of-domain generalization. Meanwhile, *ABDSynth* is slightly less accurate than the other methods, but presents an alternative in scenarios where annotated MRI data is scarce. The data and evaluation code are available for future benchmarking `https://github.com/deepakri201/AbdoBench`.

II. MATERIALS AND METHODS

A. Benchmarked methods and associated training data

1) MRSegmentator [1]: is a method based on nnU-Net [24], and is trained on multiple datasets of various sequences and modalities, including T1-weighted (T1), T2-weighted (T2), and CT scans (Table I). Training volumes are annotated using an iterative process. First, image-to-image translation is used to convert MRI volumes into pseudo-CT scans. Then, these are segmented with TotalSegmentator [3], and the resulting label maps are propagated to the original MRI scans. Finally, the segmentations are manually refined by a radiologist using MONAI Label [41]. After annotating 50 scans, an initial nnU-Net model is trained and iteratively refined as additional data becomes available. *MRSegmentator* can segment 40 regions[1]. We use model v1.2.0 published in August 2024 and implemented in Python 3.11.5.

2) MRISegmentator-Abdomen [5]: is also based on nnU-Net, and is trained solely on T1 scans. Similarly to *MRSegmentator*, a cross-modality approach is first used to convert the MRI volumes to synthetic CT scans [42] in order to obtain labels with TotalSegmentator. Once the labels are propagated

[1] `https://github.com/hhaentze/MRSegmentator`

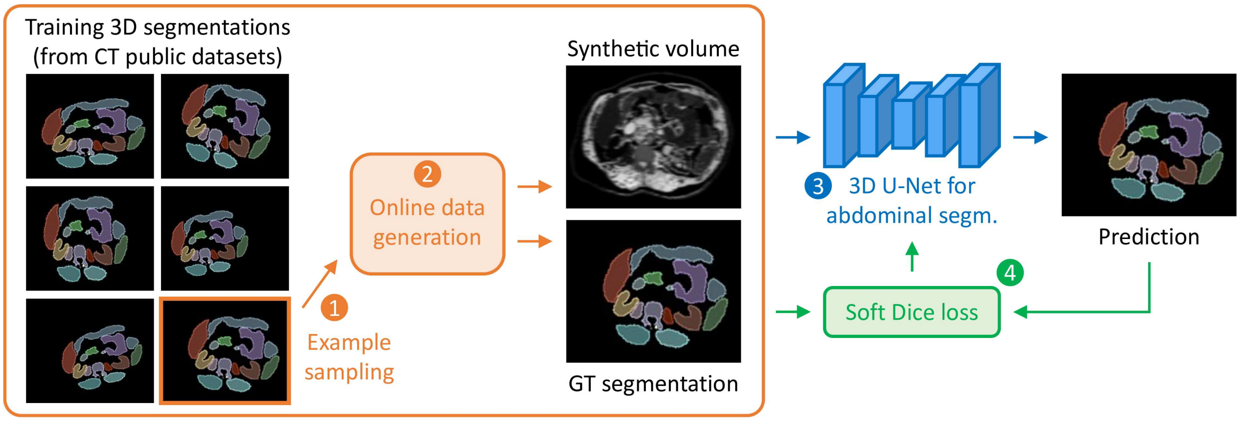


Fig. 1. Overview of an *ABDSynth* training step. 1) A CT segmentation is sampled from the training set. 2) A synthetic volume is generated using a segmentation-conditioned GMM with randomized parameters. 3,4) Abdominal volume/segmentation pairs are used to train a supervised 3D U-Net.

TABLE II
SUMMARY OF THE PUBLICLY AVAILABLE DATASETS USED FOR EVALUATION. BRACKETS DENOTE RANGES.

Dataset	# subj.	Manufacturer	Sequence	Presence of abnormalities	Regions with expert annotations	Resolution (mm)	Dimension
AMOS [35]	60	Philips	MRI (sequences not provided)	Abdominal cancer and abnormalities	Spleen, kidneys, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, adrenal glands, duodenum, bladder, prostate/uterus	[0.69×0.69×0.82] [1.95×3.00×3.00]	[192×60×64] [576×468×512]
CHAOS [36]–[38]	20	Philips	T1 dual in-phase T1 dual out-phase T2 SPIR	Healthy	Liver, spleen, right kidney, left kidney	[1.36×1.36×5.49] [2.03×2.03×9.00]	[256×256×26] [320×320×50]
LiverHCCSeg [39], [40]	17	GE Philips Siemens	T1 arterial phase	Hepatocellular carcinoma	Liver (two raters)	[0.74×0.74×1.75] [1.41×1.41×10.8]	[256×152×19] [512×512×131]

back to the MRI volumes, an iterative annotation process is used to progressively train an nnU-Net model. In total, *MRISegmentator-Abdomen* can segment 62 regions[2]. We use model v1.0.0 (June 2024, Python 3.11.10).

3) TotalSegmentator MRI [6]: is a nnU-Net-based architecture extended from TotalSegmentator for CT segmentation [3]. Starting from manual segmentations of 10 volumes, an iterative strategy is used to train a model, similar to *MRSegmentator* and *MRISegmentator-Abdomen.* In total, 59 regions can be segmented by the model[3]. We use model v2.2.0 (May 2024, Python 3.10.13).

4) ABDSynth: alleviates the need for manual MRI annotations by leveraging CT segmentations that are already publicly available. Specifically, we use the SynthSeg framework [32] to train a domain-agnostic network for MRI abdominal segmentation (Figure 1). Synthetic data is generated using a GMM conditioned on the input training label maps. Crucially, the GMM parameters are randomly sampled from uniform distributions of very wide ranges. Moreover, to further increase the diversity of the synthetic data, we apply aggressive augmentations including: affine and non-linear spatial transforms, bias field corruption, contrast augmentation, noise injection, and modeling of various voxel resolutions. Presenting the downstream segmentation network with such data forces it to learn features that are robust against these variations, such that it can segment test scans of any domain without retraining.

We train *ABDSynth* using 128 segmentations from the training set of TotalSegmentator CT [3]. These label maps are center-cropped/padded to a 300×300×250 size at 1.5mm isotropic resolution. Additional preprocessing details are given in the Appendix. In total, *ABDSynth* segments 33 regions[4]. We use the same architecture and generation parameters as in [32], and train the network for 500,000 iterations with a soft Dice loss [43]. Training takes two weeks (Nvidia A100 40GB GPU) using resources provisioned by Jetstream2 [44], [45].

B. Evaluation datasets

Our benchmark utilizes three public MRI datasets (Table II). These datasets are not used by any of the benchmarked methods for training. Overall, this cohort spans three manufacturers, five MRI sequences, healthy and diseased subjects, and a wide range of resolutions and fields-of-view, all of which enable a comprehensive evaluation for out-of-the-box deployment.

1) AMOS [35]: includes MRI scans of diseased patients with abdominal cancer and other abnormalities acquired at two centers and with eight scanners. Here, we combine the provided training and validation sets (initially used for a Grand Challenge[5]) into a single test set of 60 volumes. Annotations are provided for 15 abdominal organs (Table II). These have

[2] https://github.com/rsummers11/MRISegmentator
[3] https://github.com/wasserth/TotalSegmentator
[4] https://github.com/deepakri201/AbdoBench
[5] https://amos22.grand-challenge.org/

TABLE III

MEAN (STANDARD DEVIATIONS) DICE AND HD95 SCORES OBTAINED BY ALL BENCHMARKED APPROACHES. METHODS THAT ARE NOT TRAINED ON THE TESTED MRI SEQUENCE ARE IN ITALICS. BEST-PERFORMING METHODS ARE IN BOLD, AND ASTERISKS DENOTE STATISTICAL SIGNIFICANCE WITH ALL OTHER MODELS (5% LEVEL, BONFERRONI-CORRECTED WILCOXON SIGNED-RANK TEST).

Dataset	Region	MRSegmentator		MRISegmentator-Abdomen		TotalSegmentator MRI		ABDSynth	
		Dice	HD95 (mm)	Dice	HD95	Dice	HD95	Dice	HD95
AMOS	Liver	0.96 (0.01)	__2.87__* (1.35)	__0.97__* (0.02)	18.86 (40.15)	0.93 (0.01)	4.92 (2.35)	*0.95 (0.02)*	*4.50 (4.80)*
	Spleen	0.94 (0.03)	__2.04__* (0.98)	__0.96__* (0.04)	16.60 (60.13)	0.90 (0.02)	2.77 (0.96)	*0.94 (0.05)*	*2.52 (2.94)*
	Kidney, left	__0.95__* (0.02)	__1.99__* (0.66)	0.95 (0.10)	7.42 (24.74)	0.91 (0.07)	2.73 (1.90)	*0.92 (0.06)*	*2.51 (1.33)*
	Kidney, right	__0.95__* (0.03)	__2.25__ (1.45)	0.94 (0.06)	31.10 (58.37)	0.92 (0.12)	3.33 (7.66)	*0.92 (0.13)*	*2.83 (4.98)*
	Pancreas	0.80 (0.12)	__5.20__* (4.95)	__0.85__* (0.11)	5.52 (16.54)	0.75 (0.13)	6.67 (5.30)	*0.72 (0.17)*	*12.43 (17.88)*
	Stomach	0.87 (0.10)	__5.25__* (6.42)	__0.88__* (0.13)	11.35 (36.02)	0.86 (0.10)	5.86 (7.76)	*0.83 (0.14)*	*8.83 (10.66)*
	Gallbladder	0.78 (0.18)	5.78 (6.63)	__0.82__* (0.18)	7.40 (20.30)	0.79 (0.19)	__5.54__* (7.69)	*0.70 (0.32)*	*7.65 (8.89)*
	Duodenum	0.60 (0.16)	11.18 (7.46)	__0.69__* (0.14)	__9.02__* (5.88)	0.57 (0.17)	12.97 (9.97)	*0.56 (0.22)*	*19.96 (18.32)*
	Adrenal gland, left	0.53 (0.21)	6.39 (5.25)	__0.62__* (0.19)	__5.39__* (5.64)	0.51 (0.20)	6.48 (6.27)	*0.45 (0.24)*	*9.12 (7.92)*
	Adrenal gland, right	0.54 (0.14)	5.88 (3.97)	__0.61__* (0.13)	__3.80__* (2.54)	0.52 (0.14)	6.04 (3.95)	*0.50 (0.13)*	*6.02 (4.19)*
CHAOS T1 in-phase	Liver	__0.93__* (0.01)	__2.05__* (0.41)	*0.81 (0.12)*	*40.86 (25.42)*	0.90 (0.02)	5.28 (4.45)	*0.91 (0.03)*	*4.86 (5.76)*
	Spleen	__0.89__* (0.02)	__1.82__* (0.63)	*0.34 (0.23)*	*20.67 (5.64)*	0.87 (0.02)	2.01 (0.49)	*0.84 (0.08)*	*3.38 (2.26)*
	Kidney, left	__0.88__* (0.03)	__2.29__* (0.49)	*0.63 (0.26)*	*6.83 (9.25)*	0.79 (0.03)	3.30 (0.71)	*0.68 (0.30)*	*8.03 (8.93)*
	Kidney, right	__0.90__* (0.04)	__1.86__* (0.52)	*0.78 (0.13)*	*18.01 (32.49)*	0.80 (0.05)	3.12 (0.50)	*0.68 (0.33)*	*6.45 (6.81)*
CHAOS T1 out-phase	Liver	__0.93__* (0.01)	__2.18__* (0.50)	*0.83 (0.10)*	*30.93 (30.28)*	0.89 (0.02)	4.76 (4.22)	*0.90 (0.03)*	*4.87 (6.09)*
	Spleen	__0.88__* (0.03)	__2.11__* (1.12)	*0.45 (0.18)*	*38.88 (39.16)*	0.85 (0.03)	2.29 (0.77)	*0.76 (0.27)*	*4.25 (5.22)*
	Kidney, left	__0.87__* (0.04)	__2.44__* (0.58)	*0.77 (0.06)*	*5.48 (5.82)*	0.77 (0.04)	3.52 (0.75)	*0.74 (0.17)*	*4.18 (2.14)*
	Kidney, right	__0.88__* (0.04)	__2.06__* (0.47)	*0.74 (0.18)*	*10.21 (20.21)*	0.80 (0.03)	2.95 (0.57)	*0.68 (0.26)*	*4.71 (3.24)*
CHAOS T2 SPIR	Liver	__0.91__ (0.02)	__2.94__ (1.61)	*0.88 (0.12)*	*11.77 (19.06)*	0.91 (0.03)	4.59 (6.35)	*0.90 (0.05)*	*5.11 (6.03)*
	Spleen	0.88 (0.12)	2.39 (2.50)	*0.87 (0.22)*	*13.56 (28.42)*	0.86 (0.08)	5.31 (9.28)	___0.91___* *(0.04)*	___2.22___ *(1.67)*
	Kidney, left	0.91 (0.03)	__2.27__* (0.83)	___0.92___* *(0.02)*	*5.63 (14.56)*	0.88 (0.03)	2.67 (0.86)	*0.83 (0.09)*	*3.34 (0.86)*
	Kidney, right	__0.92__* (0.02)	__1.88__* (0.77)	*0.91 (0.05)*	*2.69 (2.55)*	0.90 (0.02)	2.56 (0.58)	*0.86 (0.14)*	*3.34 (2.25)*
LiverHCCSeg	Liver (Rater 1)	__0.93__* (0.03)	__4.93__* (3.94)	0.93 (0.07)	10.20 (21.90)	0.91 (0.04)	6.49 (4.66)	*0.90 (0.08)*	*11.49 (19.42)*

been obtained with an iterative process, where a model is first trained on a small set of annotated data to generate pseudo-labels, which are then refined by two radiologists.

2) CHAOS [36]–[38]: contains healthy subjects from the Dokuz Eylul University Hospital. It consists of three MRI sequences: T1 dual in-phase, T1 dual out-phase, and T2 SPIR (spectral pre-saturation inversion recovery). The T1 sequences are fat-suppressed, and T2 SPIR is designed to highlight the liver parenchyma[6]. Consensus ground truth segmentations are provided for the liver, spleen, right kidney, and left kidney using majority voting between three radiologists. 20 volumes are included for each sequence, for a total of 60 volumes.

3) LiverHCCSeg [39], [40]: includes 17 subjects with hepatocellular carcimona from TCGA-LIHC [46] and imaged with a T1 arterial phase sequence. All scans are provided with manual liver segmentations from two independent raters.

C. Evaluation metrics

We use Dice scores [47] and the 95th percentile of the Hausdorff distance (HD95) [48] as evaluation metrics. Dice quantifies the overlap between two segmented regions in [0,1], where 1 indicates perfect overlap and 0 indicates no overlap. HD95, expressed in millimeters, measures the 95th percentile of the surface distance between two segmentations.

III. RESULTS

We organize the benchmark results by datasets (Table III).

A. AMOS

Among the benchmarked methods, *MRISegmentator-Abdomen* achieves the highest Dice scores for the majority of organs (average gap of 0.035 Dice with *MRSegmentator*), except for the kidneys. However, it consistently exhibits high HD95 values and extreme outliers compared to the other methods. This is likely due to *MRISegmentator-Abdomen* producing implausible segmentations, including predictions of irrelevant regions, as seen in Figure 3. In contrast, *MRSegmentator* and *TotalSegmentator MRI* achieve much more spatially coherent predictions, as indicated by lower HD95 values. *ABDSynth* performs competitively on high-contrast organs such as the liver, spleen, and kidneys, but underperforms on more spatially variable regions, such as the stomach and gallbladder.

More precisely, Figure 2 reveals that all four methods yield accurate segmentations of the liver, spleen, and kidneys (Dice higher than 0.9 in all cases), but substantially lower and more variable performances for other regions. This is explained by the morphological variability, small size, and high deformability of regions like the pancreas [49], [50]. Additional complexity arises for organs such as the duodenum, where peristaltic motion during imaging introduces artifacts that further degrade segmentation performance. Moreover, while smaller Dice scores are indicative of lower performances in the adrenal glands, we highlight that Dice is known to degrade faster in such small regions. This is confirmed by the fact that results are more homogeneous across regions for HD95 than Dice.

Figure 3 shows qualitative results for each method, with a focus on the pancreas, where no method achieves anatomically correct segmentation of this region. The 3D renderings reveal over-/under-segmentation (e.g., gallbladder, spleen, and kidneys) patterns across different methods.

B. CHAOS

Table III shows that *MRSegmentator* achieves the highest Dice scores (above 0.87) and lowest HD95 (below 3mm)

[6] https://chaos.grand-challenge.org/Data/

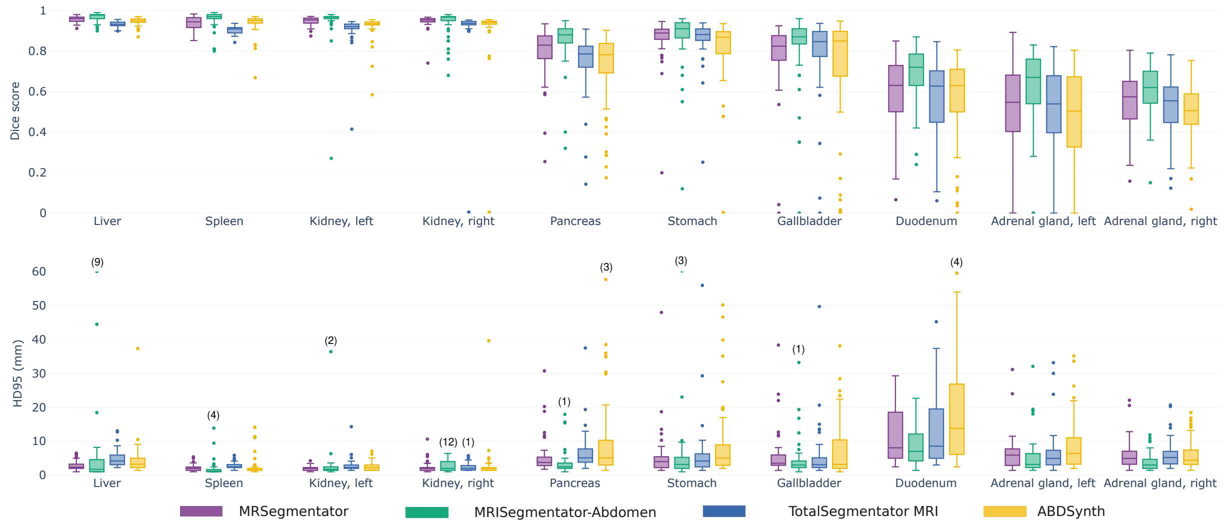
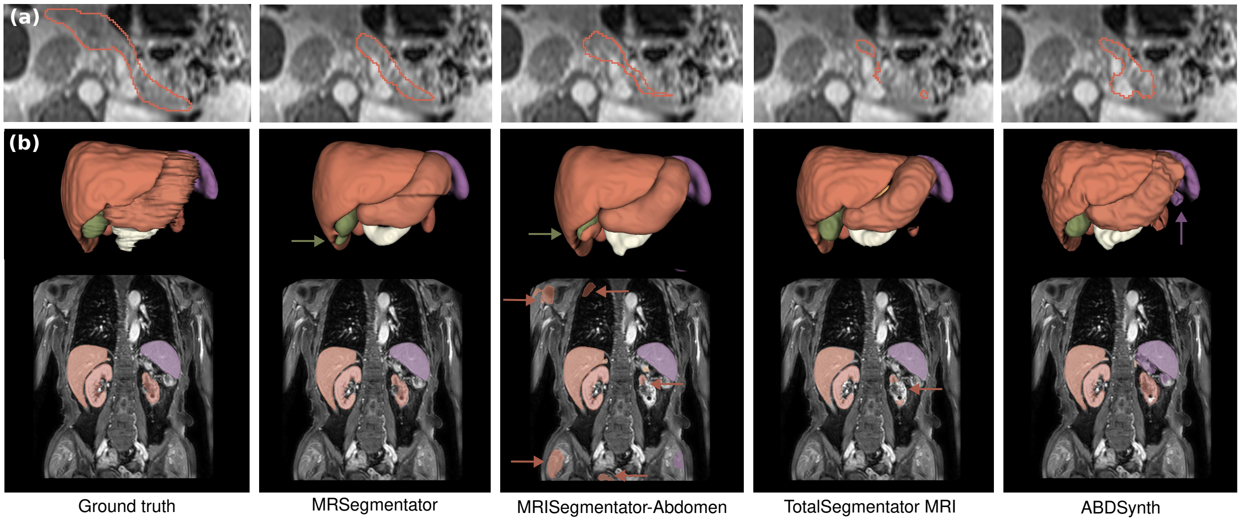


Fig. 2. Dice (top) and HD95 (bottom) boxplots for AMOS results for the four benchmarked methods. We observe similar performances for all methods across the liver, spleen, and kidneys, but highly variable results across the regions that are smaller and/or with more variable morphologies.



Ground truth MRSegmentator MRISegmentator-Abdomen TotalSegmentator MRI ABDSynth

Fig. 3. Sample segmentations by all methods on AMOS. (a) Pancreas slice, where all methods do not fully segment the region. (b) 3D renderings for another subject. Arrows indicate segmentation errors in the region of corresponding color.

among all methods and for almost all sequence types and evaluated regions. *TotalSegmentator MRI* and *ABDSynth* also yield fairly high Dice scores and low HD95 values. In contrast, *MRISegmentator-Abdomen*, which is not trained on any of the sequences used in CHAOS, displays much lower Dice scores and higher HD95 values, especially for the T1 scans.

Figure 4 illustrates the distributions of the Dice and HD95 metrics, as well as the volume repeatability across the MRI sequences used in CHAOS (T1 dual in-phase/out-phase, and T2 SPIR). In particular, we focus on the liver and right kidney, which are the most and least consistently segmented regions, respectively, among the four available labels in CHAOS. We also show qualitative segmentation examples obtained by all methods for the liver and right kidney in Figure 5.

For the liver, Dice scores are consistently high across all methods and scan types, with medians above 0.8 in all cases. However, this high overall accuracy is nuanced by the HD95 metric, for which substantial variations in standard deviations indicate local under- and over-segmentations (Figure 5). This effect is particularly visible for *MRISegmentator-Abdomen* (HD95 standard deviations above 19mm for all sequences), where typical segmentation mistakes are illustrated in Figure 5. More precisely, *MRISegmentator-Abdomen*, which has not been trained on any of the CHAOS sequences, yields an average HD95 gap of 16.6mm with the best performing method (*MRSegmentator*) across all CHAOS regions and sequences, which is far worse than the other methods. This issue is also highlighted by the volume analysis (Figure 4), where *MRISegmentator-Abdomen* displays large volumetric intra-subject differences across sequences. In comparison, the other approaches exhibit consistent liver volume estimates across sequences for most subjects, thus highlighting the accuracy of their liver segmentation predictions.

Meanwhile, we observe the opposite trend for the right

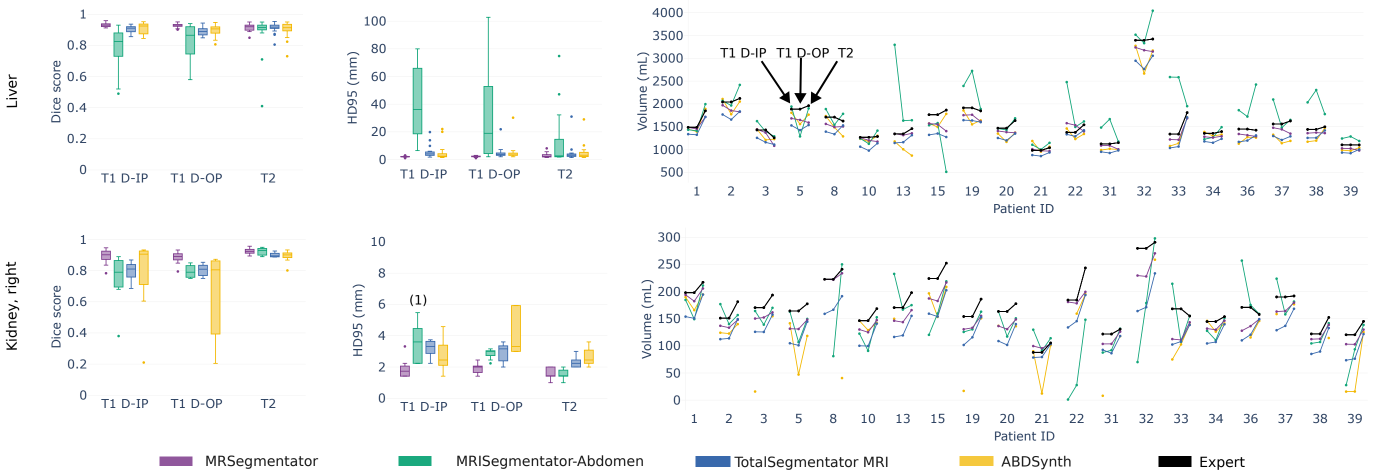


Fig. 4. Dice score (left), HD95 (middle), and volume repeatability (right) obtained on CHAOS for two representative regions (liver and right kidney) across different sequences. In the volume repeatability subfigure, the consecutive points represent T1 dual in-phase, T1 dual out-phase, and T2 SPIR, respectively. In general, the liver is more consistently segmented across MRI sequences than the right kidney.

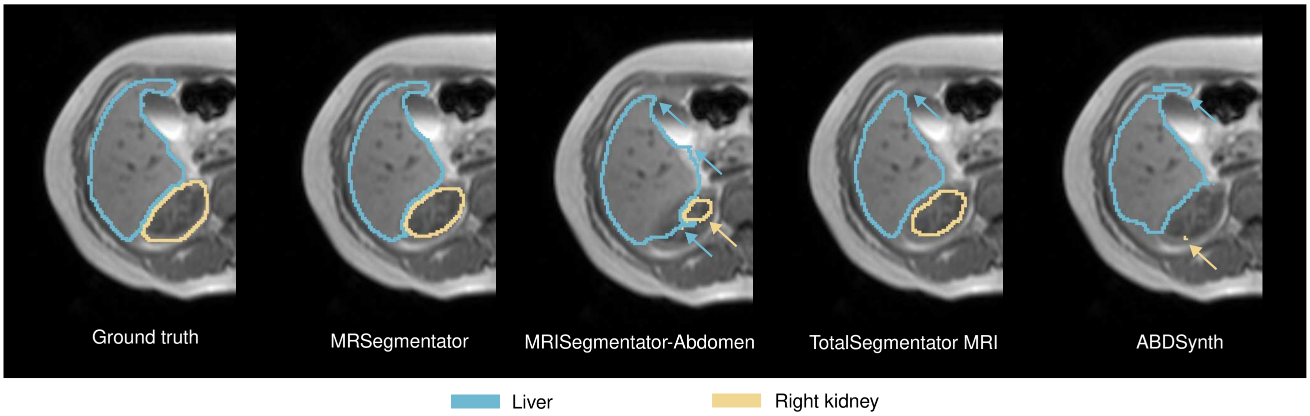


Fig. 5. CHAOS subject 39, where blue = liver and yellow = right kidney. Blue and yellow arrows point at major differences between ground truth and automated segmentations for the liver and right kidney, respectively.

kidney, where all methods obtain tighter HD95 distributions but more variable Dice scores (Figure 4). Here, the lower Dice scores are due to the substantially smaller size of the kidney, since the Dice metric is more sensitive to segmentation mistakes in smaller regions. Yet, the Dice results also reflect frequent instances of under-segmentation of the right kidney by all methods (Figure 5). In particular, we observe that *ABDSynth* fails to produce segmentations for several subjects, as indicated by missing points in the volume repeatability plot in Figure 4. Nevertheless, these segmentation mistakes remain relatively smaller compared to the liver (all methods produce substantially lower HD95 scores for the right kidney across all sequences), which may be due to the good tissue contrast with the surrounding organs.

C. LiverHCCSeg

For the LiverHCCSeg dataset, Table III presents the performance of automated methods relative to Rater 1. We observe consistently high Dice scores across all methods, ranging from 0.9 (*ABDSynth*) to 0.93 (*MRSegmentator* and *MRISegmentator-Abdomen*). In contrast, the HD95 values exhibit substantial variability, with *ABDSynth* and *MRISegmentator-Abdomen* reporting the highest means at 11.49 mm and 10.20 mm, respectively. Figure 6 illustrates this high segmentation variability across methods for a representative subject. It can be seen that while most algorithms perform well in the mid-transverse slice, discrepancies become apparent in the superior slice, where predictions vary widely. In the inferior slice, the presence of a hepatocellular carcinoma appears to degrade segmentation quality across all methods.

Since LiverHCCSeg provides liver annotations from two experts, we now compare the results of all methods against inter-rater reproducibility scores. First, the two experts show a strong overall consistency with a mean Dice score of 0.95 [39]. Remarkably, all methods yield results that are relatively close, thus highlighting the quality of the produced segmentations. The inter-rater HD is only 15.7mm [39], which is worse than any automated method. Beyond further emphasizing the good performance of the benchmarked methods, this result highlights the inter-rater reproducibility issues in annotating regions, especially for diseased tissues such as hepatocellular carcinoma in this example.

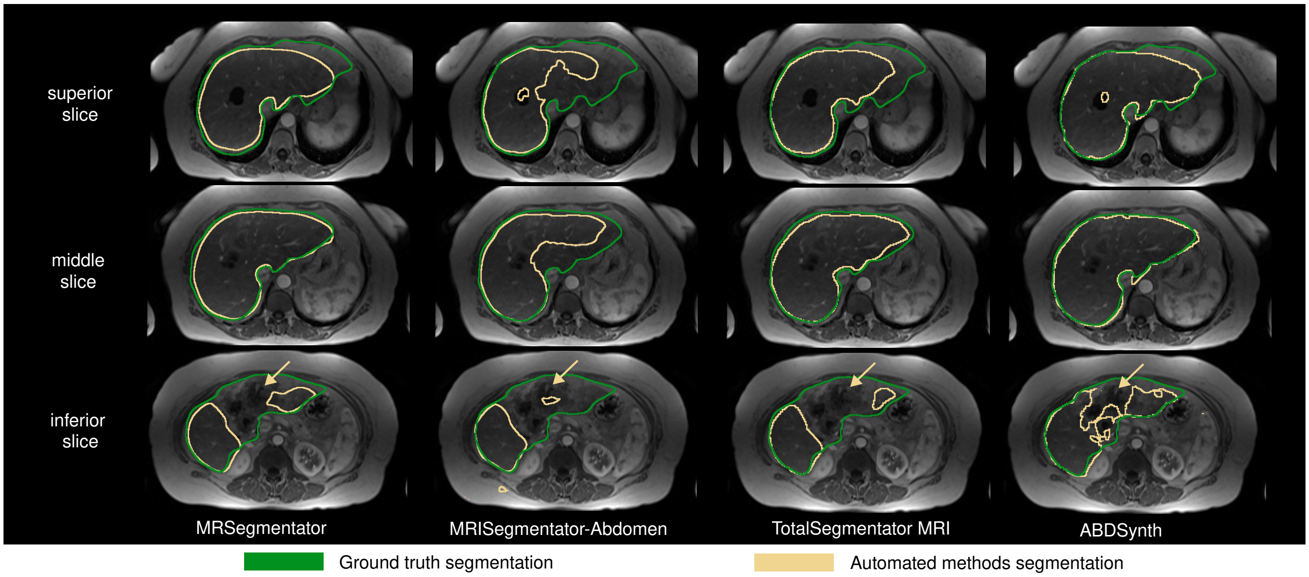


Fig. 6. Example of liver segmentations for the superior, middle, and inferior axial slices of a representative LiverHCCSeg subject. For the inferior axial slice, all methods do not segment the liver well, likely due to the presence of a hepatocellular carcinoma, as indicated by the yellow arrows.

TABLE IV
COMPARISON OF INFERENCE TIME AND MODEL SIZE.

Benchmarked method	Inference time (s)	Trainable params.
MRSegmentator	57.95 ± 53.15	31M
MRISegmentator-Abdomen	98.19 ± 69.83	31M
TotalSegmentator MRI	39.60 ± 10.34	31M
ABDSynth	21.17 ± 19.30	13M

D. Computational requirements and inference time

We now compare all methods in terms of inference time (computed on the same A100 Nvidia GPU as before) and model size (Table IV). Inference time differences among the nnU-Net-based models are explained by their preprocessing strategies, and especially by the size of the patches used for sliding-window inference: *MRSegmentator* ($96\times128\times160$ patches), *MRISegmentator-Abdomen* ($48 \times 160 \times 192$), and *TotalSegmentator MRI* ($112\times128\times160$). In contrast, *ABDSynth* represents an alternative in time-constrained scenarios, as it is faster (it does not use a patch-based strategy) and is two-thirds smaller in terms of number of parameters.

IV. DISCUSSION

In this paper, we present a thorough benchmarking of the state-of-the-art methods in MRI abdominal segmentation: *MRSegmentator*, *MRISegmentator-Abdomen*, and *TotalSegmentator MRI*. Since these methods are trained on MRI segmentations obtained by a labor-intensive iterative process involving several rounds of corrections, we also test another method *ABDSynth* (extending the SynthSeg framework) that only requires widely available CT segmentations to be trained. We perform benchmarking on a collection of three publicly available datasets, AMOS, CHAOS, and LiverHCCSeg, which cover three manufacturers, five different MRI sequences, different subject conditions (healthy and diseased patients), as well as a large range of resolutions and fields-of-view.

A. Robustness of the methods

1) Effect of sequence type: In order to analyze robustness to different sequences, we focus on the results obtained on CHAOS, which is the only evaluation dataset with multiple sequences for all subjects. Table III shows that *MRSegmentator* has the highest performance for all sequences. This can be explained by the fact that *MRSegmentator* is trained on the most diverse dataset with multiple T1 and T2 sequences (Table II). In comparison, *TotalSegmentator MRI* yields slightly lower performances, which may be due to its less abundant training data (1561 fewer scans than *MRSegmentator*). Regarding *MRISegmentator-Abdomen*, it produces lower quality segmentations on the CHAOS and LiverHCCSeg datasets, which may be due to its relatively high training resolutions compared to the CHAOS resolutions (mean in-slice resolution of 1.28mm vs. 1.62mm, and mean slice spacing of 3.1mm vs. 5mm). Finally, despite *ABDSynth* having never seen real images during training, it can accurately segment them during testing. However, *ABDSynth* fails in some cases since it does not have access to real intensity distributions during training, an issue known as the reality gap [51].

To further study robustness across sequences, we perform statistical tests between the Dice scores obtained by each method on each region of CHAOS by using the Friedman chi-square test with a significance level of 0.05. Almost all of the regions show statistically significant results, except in three cases (*ABDSynth* for liver and spleen, *MRSegmentator* for spleen, and *TotalSegmentator MRI* for spleen), thus emphasizing the remaining performance differences across sequences.

2) Presence of pathologies: While the CHAOS dataset comprises only healthy subjects, the AMOS and LiverHCCSeg cohorts contain patients with cancer and other abnormalities. Overall, even though the benchmarked methods are trained on datasets containing various abnormalities (e.g., tumors, cysts, etc.), we observe that the evaluated methods may lack robustness to diverse pathological conditions. Importantly, we

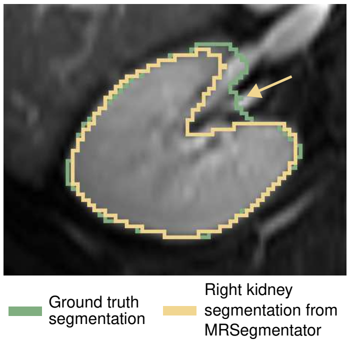

🟩 Ground truth segmentation
🟨 Right kidney segmentation from MRSegmentator

Fig. 7. Right kidney segmentations for CHAOS subject 1. The yellow arrow points to major differences between ground truth and automated segmentations, where the ground truth includes the renal pelvis.

note that the public datasets used here for benchmarking do not provide subject-specific medical information. Therefore, it is challenging to discern whether a model's lower performance stems from subject-specific abnormalities or from the inherent difficulty of segmenting certain anatomical regions. This ambiguity is illustrated in Figure 3, which shows consistently poor pancreas segmentation across all methods for a patient with liver pathology, which is likely exacerbated by the pancreas being inherently difficult to segment. Interestingly, for the AMOS dataset, high segmentation performance is observed across all methods for the primary abdominal organs, liver, spleen, and kidneys, as shown in Table III, despite the presence of pathologies. In contrast, the LiverHCCSeg dataset reveals more variable segmentation quality, likely due to the presence of hepatocellular carcinoma. As illustrated in Figure 6, although all methods achieve high Dice scores (Table III), segmentation accuracy tends to degrade in inferior slices affected by this pathology, while segmentations of mid-axial slices remain comparatively consistent.

B. Inconsistencies in segmentation conventions

Figure 7 illustrates differences in anatomical conventions between the CHAOS ground truth annotations and those used in the training data of the evaluated methods. Here, the expert reference includes the renal pelvis as part of the kidney segmentation, whereas the automated methods exclude this region (yellow arrow). These discrepancies in the definition of anatomical boundaries contribute to the lower volume estimates observed in the volume repeatability analysis (Figure 4). This example shows a limitation of our study, where the evaluated methods and benchmarking datasets might use different conventions for some of the regions. More generally, this highlights the importance of identifying semantic inconsistencies in annotation protocols when deploying models across heterogeneous datasets.

C. Key difference between benchmarked methods

Three of the benchmarked methods, *MRSegmentator*, *MRISegmentator-Abdomen*, and *TotalSegmentator MRI* require expert involvement during training, where a clinician or radiologist guides the model through an iterative labels refinement process. In addition to this labor-intensive training paradigm, these methods rely on large datasets: *MRSegmentator* was trained on 2,649 MRI and CT volumes, *MRISegmentator-Abdomen* on 780 MRI volumes, and *TotalSegmentator MRI* on 1,088 MRI and CT volumes. CT data was utilized by *MRSegmentator* and *TotalSegmentator MRI* to improve robustness and cross-modality segmentation capabilities. Furthermore, because these models are trained on specific sequences, their performances degrade when applied to unseen sequences. This limitation is particularly evident in the performance of *MRISegmentator-Abdomen* on the CHAOS dataset (Table III), where the model shows poorer results across all three scan types, due to a lack of training on those sequences. Consequently, adapting these methods to segment a new MRI sequence would require additional retraining or fine-tuning.

In contrast, *ABDSynth* requires only a single set of annotated CTs, which are widely available, for training. This represents a significant advantage given the relative scarcity of large, annotated MRI datasets compared to CT. By leveraging annotated CT data for MRI segmentation, *ABDSynth* substantially reduces the burden of manual labeling in the MRI imaging space. Furthermore, the method's synthetic data generation approach enables adaptation to new MRI sequence types without requiring additional expert-annotated MRI datasets and retraining.

However, as shown in Table III, *ABDSynth* generally yields lower segmentation performance compared to the other benchmarked methods. Moreover, there are multiple instances where *ABDSynth* produces very poor segmentation outputs, most notably for the T1 out-phase scans in the CHAOS dataset, as reflected by the missing volumes in the volume repeatability plot in Figure 4. These observations highlight an inherent trade-off among the benchmarked methods, balancing *(i)* training and annotation effort, *(ii)* the diversity of sequences used for training, and *(iii)* overall segmentation performance.

V. Conclusion

We presented a benchmarking study of abdominal MRI segmentation methods, including three state-of-the-art models trained on real data and one method trained on synthetically generated data. The models are evaluated on publicly available datasets from multiple grand challenges, as well as a multi-rater liver segmentation dataset. Among the evaluated methods, *MRISegmentator-Abdomen* achieved high Dice scores on the AMOS dataset but exhibited high HD95 values, indicating many outlier segments. Moreover, its performance on the CHAOS dataset was notably lower, likely due to the presence of MRI sequences not included in its training set. In contrast, *MRSegmentator* demonstrated consistent performance across all datasets, with moderately high Dice scores and lower variability, suggesting greater robustness, which is potentially attributed to its more diverse training set.

Our benchmarking approach suffers from several limitations, which we plan to address in future work. First, while this study focused on a core set of representative automated methods, we did not evaluate all other available tools, and notably TotalVibeSegmentator [52], which is specifically designed for segmenting volumetric interpolated breath-hold examination (VIBE) sequences. Inclusion of such specialized models may be considered in future evaluations targeting sequence-specific performance. Secondly, a pathology-specific performance analysis was not feasible due to dataset constraints, as the AMOS dataset lacks pathology labels, CHAOS includes only healthy subjects, and although LiverHCCSeg contains patients who all have hepatocellular carcinoma, it is limited by the small number of subjects in the dataset (17). Future benchmarking efforts would benefit from larger, more diverse datasets with well-annotated pathological labels to enable an evaluation of segmentation performance as a function of different pathologies.

Overall, by releasing our evaluation code as well as the diverse cohort of testing MRI scans, the proposed benchmark represents a first step towards precise and thorough benchmarking of current and future methods for MRI multi-organ abdominal segmentation, a rapidly evolving and promising field for clinical practice.

APPENDIX

Preprocessing the training data of ABDSynth

In SynthSeg [32], synthetic data is generated using a GMM conditioned on training label maps, where each anatomical label is associated with a single Gaussian distribution. While effective, representing the intensities of a given label by a single Gaussian can be insufficient for labels that include heterogeneous substructures with distinct intensity profiles. For example, regions like the renal cortex and medulla in the kidneys, or hepatic vasculature, contain fine-grained differences that are not well captured by a single Gaussian. To address this, we refine the label maps used for synthetic data generation to introduce finer anatomical detail. We adopt a similar strategy to Billot et al. when they extended SynthSeg to cardiac segmentation [32]. Using the original TotalSegmentator CT scans, we subdivide each label into subregions by clustering the corresponding intensities using expectation-maximization (EM) [53]. In order to capture different levels of granularity, we randomly sample the number of clusters from $\{1, 2, 3\}$ for each foreground label. We also apply the same strategy to the background class, but we sample the number of clusters in $\{3, 4, 5, 6, 7\}$ to account for the greater variability of the underlying tissues (Algorithm 1). Clustering is performed dynamically during synthetic volume generation, with the number of clusters selected at runtime.

After subdividing labels into substructures, we also simulate different scanning poses that are more specific to MRI acquisitions, and especially poses where only the trunk of the subject is acquired. This is achieved by removing the arms of the subject with a 0.5 probability, where the arm regions have been defined using the 3D Slicer Sandbox extension[7].

Algorithm 1 Background and foreground labels clustering used for synthetic image generation

```
for $K_{BG} \in \texttt{Rand}\{3, 4, 5, 6, 7\}$ do
    $P(x) = \sum_{k=1}^{K_{BG}} \mathcal{N}(x \mid \mu_k, \sigma_k^2)\pi_k$
    Optimize $\theta_{BG} = \{\mu_k, \sigma_k^2, \pi_k \mid k = 1, \ldots, K_{BG}\}$ using Expectation-Maximization (EM) algorithm
end for
for label in $[1, \ldots, N_{seg}]$ do
    for $K_{FG} \in \texttt{Rand}\{1, 2, 3\}$ do
        $P(x_{\text{label}}) = \sum_{k=1}^{K_{FG}} \mathcal{N}(x_{\text{label}} \mid \mu_k, \sigma_k^2)\pi_k$
        Optimize $\theta_{FG} = \{\mu_k, \sigma_k^2, \pi_k \mid k = 1, \ldots, K_{FG}\}$ using EM algorithm
    end for
end for
$N_{\text{Seg}}$ refers to the total number of labels used to annotate a specific CT scan.
$\pi_k$ refers to the weight of the *k-th* Gaussian component.
```

REFERENCES

[1] H. Häntze, L. Xu, F. Dorfner, L. Donle, D. Truhn, H. Aerts, M. Prokop, B. van Ginneken, A. Hering, L. Adams, and K. Bressem, "MRSegmentator: Robust multi-modality segmentation of 40 classes in MRI and CT sequences," *arXiv:2405.06463*, 2024.

[2] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, and B. Liu, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.

[3] J. Wasserthal, H. Breit, M. Meyer, M. Pradella, D. Hinck, A. Sauter, T. Heye, D. Boll, J. Cyriac, S. Yang, and M. Bach, "TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. 102789, 2023.

[4] J. Wasserthal, "Dataset with segmentations of 104 important anatomical structures in 1204 CT images," 2023. [Online]. Available: https://zenodo.org/records/6802614

[5] Y. Zhuang, T. Mathai, P. Mukherjee, B. Khoury, B. Kim, B. Hou, N. Rabbee, and R. Summers, "MRISegmentator-Abdomen: A fully automated multi-organ and structure segmentation tool for T1-weighted abdominal MRI," *arXiv.2405.05944*, 2024.

[6] T. D'Antonoli, L. Berger, A. Indrakanti, N. Vishwanathan, J. Weiß, M. Jung, Z. Berkarda, A. Rau, M. Reisert, T. Küstner, and A. Walter, "TotalSegmentator MRI: Sequence-independent segmentation of 59 anatomical structures in MR images," *arXiv:2405.19492*, 2024.

[7] A. Fedorov, W. Longabaugh, D. Pot, D. Clunie, S. Pieper, D. Gibbs, C. Bridge, M. Herrmann, A. Homeyer, R. Lewis, and H. Aerts, "National cancer institute imaging data commons: Toward transparency, reproducibility, and scalability in imaging artificial intelligence," *Radiographics*, vol. 43, no. 12, p. e230180, 2023.

[8] L. Lenchik, L. Heacock, A. Weaver, R. Boutin, T. Cook, J. Itri, C. Filippi, R. Gullapalli, J. Lee, M. Zagurovskaya, and T. Retson, "Automated segmentation of tissues using CT and MRI: a systematic review," *Academic radiology*, vol. 26, no. 12, pp. 1695–706, 2019.

[9] S. DeSouza, R. Singh, H. Yoon, R. Murphy, L. Plank, and M. Petrov, "Pancreas volume in health and disease: a systematic review and meta-analysis," *Expert review of gastroenterology & hepatology*, vol. 12, no. 8, pp. 757–66, 2018.

[10] F. Zöllner, E. Svarstad, A. Munthe-Kaas, L. Schad, A. Lundervold, and J. Rørvik, "Assessment of kidney volumes from mri: acquisition and segmentation techniques," *American Journal of Roentgenology*, vol. 199, no. 5, pp. 1060–9, 2012.

[11] P. Dirix, K. Haustermans, and V. Vandecaveye, "The value of magnetic resonance imaging for radiotherapy planning," *In Seminars in radiation oncology*, vol. 24, no. 3, pp. 151–159, 2014.

[12] P. Keall, C. Brighi, C. Glide-Hurst, G. Liney, P. Liu, S. Lydiard, C. Paganelli, T. Pham, S. Shan, A. Tree, and U. van der Heide, "Integrated MRI-guided radiotherapy—opportunities and challenges," *Nature Reviews Clinical Oncology*, vol. 19, no. 7, pp. 458–70, 2022.

[13] R. Gillies, P. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–77, 2016.

[14] M. Barat, A. Pellat, C. Hoeffel, A. Dohan, R. Coriat, E. Fishman, S. Nougaret, L. Chu, and P. Soyer, "CT and MRI of abdominal cancers: current trends and perspectives in the era of radiomics and artificial

[7] https://github.com/PerkLab/SlicerSandbox#remove-ct-table

intelligence," *Japanese journal of radiology*, vol. 42, no. 3, pp. 246–60, 2024.

[15] H. Heerkens, W. Hall, X. Li, P. Knechtges, E. Dalah, E. Paulson, C. van den Berg, G. Meijer, E. Koay, C. Crane, and K. Aitken, "Recommendations for MRI-based contouring of gross tumor volume and organs at risk for radiation therapy of pancreatic cancer," *Practical radiation oncology*, vol. 7, no. 2, pp. 126–36, 2017.

[16] G. Podobnik, B. Ibragimov, P. Peterlin, P. Strojan, and T. Vrtovec, "vOARiability: Interobserver and intermodality variability analysis in oar contouring from head and neck CT and MR images," *Medical Physics*, vol. 51, no. 3, pp. 2175–86, 2024.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, Proceedings, Part III 18*, pp. 234–241, 2015.

[18] B. Billot, K. Greve, D.and Van Leemput, B. Fischl, J. Iglesias, and A. Dalca, "A learning strategy for contrast-agnostic mri segmentation," *Medical Imaging with Deep Learning (MIDL)*, p. 175–93, 2020.

[19] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, p. 45–59, 2010.

[20] Y. Chen, D. Ruan, J. Xiao, L. Wang, B. Sun, R. Saouaf, W. Yang, D. Li, and Z. Fan, "Fully automated multiorgan segmentation in abdominal magnetic resonance imaging with deep neural networks," *Medical physics*, vol. 47, no. 10, pp. 4971–82, 2020.

[21] T. Kart, M. Fischer, T. Küstner, T. Hepp, F. Bamberg, S. Winzeck, B. Glocker, D. Rueckert, and S. Gatidis, "Deep learning-based automated abdominal organ segmentation in the uk biobank and german national cohort magnetic resonance imaging studies," *Investigative Radiology*, vol. 56, no. 6, pp. 401–8, 2021.

[22] A. Rickmann, J. Senapati, O. Kovalenko, A. Peters, F. Bamberg, and C. Wachinger, "Abdomennet: deep neural network for abdominal organ segmentation in epidemiologic imaging studies," *BMC medical imaging*, vol. 22, no. 1, p. 168, 2022.

[23] A. Amjad, J. Xu, D. Thill, Y. Zhang, J. Ding, E. Paulson, W. Hall, B. Erickson, and X. Li, "Deep learning auto-segmentation on multi-sequence magnetic resonance images for upper abdominal organs," *Frontiers in oncology*, vol. 13, pp. 1–12, 2023.

[24] F. Isensee, P. Jaeger, S. Kohl, J. Petersen, and K. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–11, 2021.

[25] G. N. C. G. Consortium, "The german national cohort: aims, study design and organization," *European journal of epidemiology*, vol. 29, no. 5, pp. 371–82, 2014.

[26] T. Mathai, S. Lee, D. Elton, T. Shen, Y. Peng, Z. Lu, and R. Summers, "Lymph node detection in t2 mri with transformers," *In Medical imaging 2022: computer-aided diagnosis, SPIE.*, vol. 12033, pp. 869–873, 2022.

[27] T. Mathai, S. Lee, T. Shen, Z. Lu, and R. Summers, "Universal lymph node detection in t2 mri using neural networks," *International journal of computer assisted radiology and surgery*, vol. 18, no. 2, pp. 313–8, 2023.

[28] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev, "Domain shift in computer vision models for MRI data analysis: an overview," in *Thirteenth International Conference on Machine Vision*, vol. 11605. SPIE, 2021, pp. 126–133.

[29] J. Li, Q. Chen, H. Ding, H. Liu, and L. Wan, "A 3d unsupervised domain adaptation framework combining style translation and self-training for abdominal organs segmentation," in *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation*. MICCAI, 2024, pp. 209–224.

[30] J. Wu, G. Zhang, X. Qi, H. Wang, X. Liu, and G. Wang, "Unsupervised domain adaptation for abdominal organ segmentation using pseudo labels and organ attention cyclegan," in *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation*. MICCAI, 2024, pp. 225–242.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[32] B. Billot, D. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. Dalca, and J. Iglesias, "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical image analysis*, vol. 86, p. 102789, 2023.

[33] B. Billot, C. Magdamo, Y. Cheng, S. E. Arnold, S. Das, and J. E. Iglesias, "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets," *Proceedings of the National Academy of Sciences*, vol. 120, no. 9, 2023.

[34] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30, 2017.

[35] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhang, W. Ma, X. Wan, and P. Luo, "AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 722–32, 2022.

[36] A. Kavur, A. Selver, O. Dicle, M. Barış, and N. Gezer, "CHAOS - combined (CT-MR) healthy abdominal organ segmentation challenge data," 2019. [Online]. Available: https://zenodo.org/records/3431873

[37] A. Kavur, N. Gezer, M. Barış, Y. Şahin, S. Özkan, B. Baydar, U. Yüksel, c. Kılıkçıer, c. Olut, G. Akar, and G. Ünal, "Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors," *Diagnostic and Interventional Radiology*, vol. 26, no. 1, p. 11, 2020.

[38] A. Kavur, N. Gezer, M. Barış, S. Aslan, P. Conze, V. Groza, D. Pham, S. Chatterjee, P. Ernst, S. Özkan, and B. Baydar, "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[39] M. Gross, S. Arora, S. Huber, A. Kücükkaya, and J. Onofrey, "LiverHCCSeg: A publicly available multiphasic MRI dataset with liver and HCC tumor segmentations and inter-rater agreement analysis," *Data in Brief*, vol. 51, p. 109662, 2023.

[40] M. Gross, S. Arora, S. Huber, A. Kucukkaya, and J. Onofrey, "LiverHCCeg: A publicly available multiphasic MRI dataset with liver and HCCC tumor segmentations and inter-rater agreement analysis," *Zenodo*, 2023.

[41] A. Diaz-Pinto, S. Alle, V. Nath, Y. Tang, A. Ihsani, M. Asad, F. Pérez-García, P. Mehta, W. Li, M. Flores, H. Roth, T. Vercauteren, D. Xu, P. Dogra, S. Ourselian, A. Feng, and C. MJ, "Monai label: A framework for ai-assisted interactive labeling of 3D medical images," *Medical Image Analysis*, vol. 95, 2024.

[42] Y. Zhuang, T. Mathai, P. Mukherjee, and R. Summers, "Segmentation of pelvic structures in T2 MRI via MR-to-CT synthesis," *Computerized Medical Imaging and Graphics*, vol. 112, p. 102335, 2024.

[43] A. S. Milletari F, Navab N, "Fully convolutional neural networks for volumetric medical image segmentation," *In 2016 fourth international conference on 3D vision (3DV)*, pp. 65–571, 2016.

[44] D. Hancock, J. Fischer, J. Lowe, W. Snapp-Childs, M. Pierce, S. Marru, J. Coulter, M. Vaughn, B. Beck, N. Merchant, E. Skidmore, and G. Jacobs, "Jetstream2: Accelerating cloud computing via jetstream." pp. 1–8, 2021.

[45] T. Boerner, S. Deems, T. Furlani, S. Knuth, and J. Towns, "ACCESS: Advancing innovation: NSF's advanced cyberinfrastructure coordination ecosystem: Services & support," 2023.

[46] B. Erickson, S. Kirk, Y. Lee, O. Bathe, M. Kearns, C. Gerdes, K. Rieger-Christ, and J. Lemmerman, "The cancer genome atlas liver hepatocellular carcinoma collection (TCGA-LIHC) (version 5) [data set]," 2016. [Online]. Available: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=6885436

[47] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[48] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–63, 1993.

[49] C. Noel, F. Zhu, A. Lee, H. Yanle, and P. Parikh, "Segmentation precision of abdominal anatomy for mri-based radiotherapy," *Medical Dosimetry*, vol. 39, no. 3, pp. 212–7, 2014.

[50] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Pancreas segmentation in mri using graph-based decision fusion on convolutional neural networks," *In Medical Image Computing and Computer Assisted Intervention - MICCAI 2017: 20th International Conference*, pp. 674–682, 2017.

[51] N. Jakobi, P. Husbands, and I. Harvey, "Noise and the reality gap: The use of simulation in evolutionary robotics," in *European conference on artificial life*, 1995, pp. 704–720.

[52] R. Graf, P. Platzek, E. Riedel, C. Ramschütz, S. Starck, H. Möller, M. Atad, H. Völzke, R. Bülow, C. Schmidt, and J. Rüdebusch, "Totalvibesegmentator: Full body mri segmentation for the nako and uk biobank," *arXiv:2406.00125*, 2024.

[53] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.