Touchstone Benchmark

Are we on the right way for evaluating medical segmentation?

Pedro R. A. S. Bassi

Johns Hopkins University

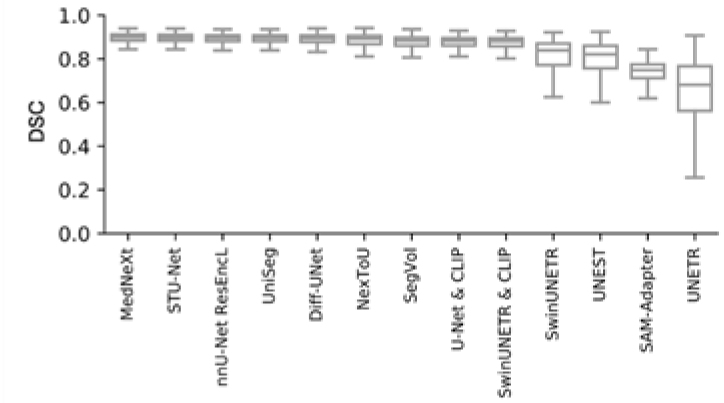
University of Bologna

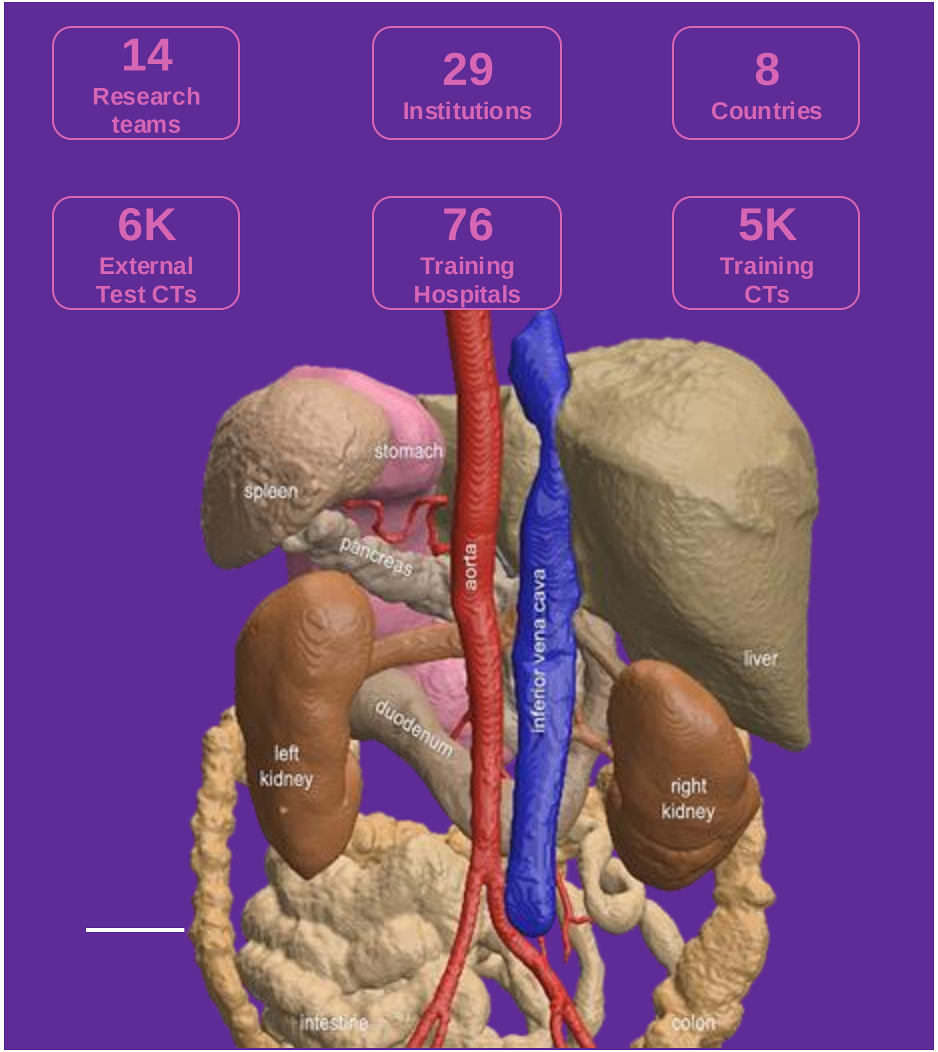
Italian Institute of Technology

E: psalvad2@jh.edu

Touchstone Ideals for AI evaluation:

- External (OOD) evaluation
- Large test set
- Analysis by age, sex, race, diagnosis, and more
- AI inventors' participation
- Long-term commitment





Scale

Table 1: **Related benchmarks & our innovations.** We compare Touchstone with influential CT segmentation benchmarks in light of the five contributions presented in the introduction.

contribution	promoting superior OOD performance with a large and diverse training dataset (#1)			boosting results' significance & large-scale OOD test (#1, #2)	multi-faceted evaluation (#3)	encouraging innovative AI (#4, #5)
benchmark	# CT scans train	# hospitals train	# countries train	# CT scans test	AI consistency analysis	targeted invitation
MSD-CT [2]	947†	1	1	465 IID	none	no
FLARE'22 [53]	2,050†	22	5+	200 IID, 600 OOD	sex, age	no
FLARE'23 [55]	4,000†	30	n/a	n/a	n/a	no
KiTS21 [29]	300	50+	1	100 OOD	sex, race	no
AMOS22-CT [38]	200	3	1	78 IID, 122 OOD	none	no
LiTS [9]	130	7	5	70 IID	none	no
BTCV [41]	30	1	1	20 IID	none	no
CHAOS-CT [71]	20	1	1	20 IID	none	no
Touchstone (ours)	**5,195**	**76**	**8**	**5,903 OOD**	**sex, age, race**	**yes**

† Partially labeled: annotations for each organ do not cover the entire dataset, and/or may contain unlabeled samples.

Touchstone Benchmark: Are We on the Right Way for Evaluating AI Algorithms for Medical Segmentation?

Pedro R. A. S. Bassi[1,2,3,*] Wenxuan Li[1,*] Yucheng Tang[4] Fabian Isensee[5,6]
Zifu Wang[7] Jieneng Chen[1] Yu-Cheng Chou[1] Yannick Kirchhoff[5,8,9]
Maximilian Rokuss[5,8] Ziyan Huang[10] Jin Ye[11] Junjun He[11] Tassilo Wald[5,6]
Constantin Ulrich[5] Michael Baumgartner[5,6] Saikat Roy[5,8]
Klaus H. Maier-Hein[5,12] Paul Jaeger[6,13] Yiwen Ye[14] Yutong Xie[15] Jianpeng Zhang[16]
Ziyang Chen[14] Yong Xia[14] Zhaohu Xing[17] Lei Zhu[17,18] Yousef Sadegheih[19]
Afshin Bozorgpour[19] Pratibha Kumari[19] Reza Azad[20] Dorit Merhof[19,21]
Pengcheng Shi[22] Ting Ma[22] Yuxin Du[23] Fan Bai[23,24] Tiejun Huang[23,25]
Bo Zhao[10,23] Haonan Wang[18] Xiaomeng Li[18] Hanxue Gu[26]
Haoyu Dong[26] Jichen Yang[26] Maciej A. Mazurowski[26] Saumya Gupta[27]
Linshan Wu[18] Jiaxin Zhuang[18] Hao Chen[28] Holger Roth[4] Daguang Xu[4]
Matthew B. Blaschko[7] Sergio Decherchi[29] Andrea Cavalli[2,29,30]
Alan L. Yuille[1†] Zongwei Zhou[1†]

[1]Department of Computer Science, Johns Hopkins University
[2]Department of Pharmacy and Biotechnology, University of Bologna
[3]Center for Biomolecular Nanotechnologies, Istituto Italiano di Tecnologia
[4]NVIDIA
[5]Division of Medical Image Computing, German Cancer Research Center (DKFZ)
[6]Helmholtz Imaging, German Cancer Research Center (DKFZ)
Full affiliations are given in Appendix F.

Code, Models & Data: https://github.com/MrGiovanni/Touchstone

Abstract

How can we test AI performance? This question seems trivial, but it isn't. Standard benchmarks often have problems such as in-distribution and small-size test sets, oversimplified metrics, unfair comparisons, and short-term outcome pressure. As a consequence, good performance on standard benchmarks does not guarantee success in real-world scenarios. To address these problems, we present Touchstone, a large-scale collaborative segmentation benchmark of 9 types of abdominal organs. This benchmark is based on 5,195 training CT scans from 76 hospitals around the world and 5,903 testing CT scans from 11 additional hospitals. This diverse test set enhances the statistical significance of benchmark results and rigorously evaluates AI algorithms across out-of-distribution scenarios. We invited 14 inventors of 19 AI algorithms to train their algorithms, while our team, as a third party, independently evaluated these algorithms. In addition, we also evaluated pre-existing AI frameworks—which, differing from algorithms, are more flexible and can support different algorithms—including MONAI from NVIDIA, nnU-Net from DKFZ, and numerous other open-source frameworks. We are committed to expanding this benchmark to encourage more innovation of AI algorithms for the medical domain.

Results

model	organization	average DSC	paper
MedNeXt	DKFZ	89.2	arXiv 2303.09975
STU-Net-B	Shanghai AI Lab	89.0	arXiv 2304.06716
MedFormer	Rutgers	89.0	arXiv 2203.00131
nnU-Net ResEncL	DKFZ	88.8	arXiv 1809.10486
UniSeg	NPU	88.8	arXiv 2304.03493
Diff-UNet	HKUST	88.5	arXiv 2303.10326
LHU-Net	UR	88.0	arXiv 2404.05102
NexToU	HIT	87.8	arXiv 2305.15911
SegVol	BAAI	87.1	arXiv 2311.13385
U-Net & CLIP	CityU	87.1	arXiv 2301.00785
Swin UNETR & CLIP	CityU	86.7	arXiv 2301.00785
Swin UNETR	NVIDIA	80.1	arXiv 2211.11537
UNesT	NVIDIA	79.1	arXiv 2303.10745
SAM-Adapter	Duke	73.4	arXiv 2404.09957
UNETR	NVIDIA	64.4	arXiv 2111.04004

Table 2: **External validation on proprietary JHH dataset** (N**=5,160).** Performance is given as DSC score (mean±s.d.). For each class, we bold the best-performing results and highlight the runners-up, which show no significant difference from the best results at $p = 0.05$ level, in red. Architectures are grouped by their frameworks and sorted in ascending order based on the number of parameters. CNNs based on the nnU-Net framework have the best performance on most classes, but other models excel at specific structures (e.g., the graph neural network-based NeXToU for aorta, and the diffusion-based Diff-UNet for kidneys). The NSD results are reported in Appendix Table 9.

framework	architecture	param	spleen	kidneyR	kidneyL	gallbladder	liver
nnU-Net	UniSeg† [83]	31.0M	94.9±6.0	92.2±7.2	91.5±7.0	84.7±12.6	96.1±4.4
	MedNeXt [64]	61.8M	95.2±6.3	92.6±7.4	91.8±7.3	85.3±12.9	96.3±4.5
	NexToU [66]	81.9M	94.7±8.1	90.1±9.5	89.6±9.3	82.3±17.0	95.7±5.5
	STU-Net-B [34]	58.3M	95.1±6.4	92.5±7.3	91.9±7.2	85.5±12.3	96.2±4.8
	STU-Net-L [34]	440.3M	95.2±6.1	92.5±7.1	91.8±7.1	85.7±11.8	96.3±4.4
	STU-Net-H [34]	1457.3M	95.2±5.9	92.6±6.9	91.9±7.1	**86.0±11.6**	96.3±4.4
	U-Net [62]	31.1M	95.1±6.3	92.7±6.9	91.9±7.2	84.7±13.1	96.2±4.5
	ResEncL [35, 37]	102.0M	95.2±6.3	92.6±7.0	91.9±6.9	84.9±13.0	96.3±4.5
	ResEncL*	102.0M	95.1±6.2	92.7±6.9	91.9±7.1	84.9±12.8	96.3±4.5
Vision-Language	U-Net & CLIP [46]	19.1M	94.3±6.9	91.9±7.8	91.1±8.8	82.1±15.4	96.0±4.3
	Swin UNETR & CLIP [46]	62.2M	94.1±7.7	91.7±9.1	91.0±9.1	80.2±18.3	95.8±5.6
MONAI	LHU-Net [65]	8.6M	94.9±6.3	92.5±7.0	91.8±7.4	83.9±14.5	96.2±4.3
	UCTransNet [72]	68.0M	90.2±11.9	86.5±14.6	86.9±12.8	77.8±19.5	93.6±6.4
	Swin UNETR [68]	72.8M	92.7±8.8	89.8±11.1	89.7±10.2	76.9±20.7	95.2±5.3
	UNesT [85]	87.2M	93.2±7.1	90.9±8.1	90.1±8.2	75.1±21.2	95.3±5.0
	UNETR [25]	101.8M	91.7±10.1	90.1±9.4	89.2±9.6	74.7±20.4	95.0±5.3
	SegVol† [18]	181.0M	94.5±6.9	92.5±7.1	91.8±7.3	79.3±18.8	96.0±4.7
n/a	SAM-Adapter† [23]	11.6M	90.5±8.8	90.4±7.9	87.3±9.6	49.4±22.9	94.1±5.3
	MedFormer [19]	38.5M	**95.5±6.1**	**92.8±7.3**	91.9±7.4	85.3±13.6	**96.4±4.4**
	Diff-UNet [81]	434.0M	95.0±6.9	92.8±7.4	**91.9±7.5**	83.8±14.8	96.2±4.7

framework	architecture	param	stomach	aorta	postcava	pancreas	average
nnU-Net	UniSeg† [83]	31.0M	93.3±6.0	82.3±10.3	81.2±8.1	82.7±10.4	88.8±5.0
	MedNeXt [64]	61.8M	93.5±6.0	83.1±10.2	81.3±8.3	83.3±11.0	**89.2±5.1**
	NexToU [66]	81.9M	92.7±7.5	86.4±8.7	78.1±9.1	80.2±13.5	87.8±6.2
	STU-Net-B [34]	58.3M	93.5±6.0	82.1±10.5	**81.3±8.2**	83.2±10.7	89.1±5.3
	STU-Net-L [34]	440.3M	93.7±5.6	81.0±10.9	81.3±8.2	83.4±10.7	89.0±5.0
	STU-Net-H [34]	1457.3M	**93.7±5.7**	81.1±10.9	81.1±8.2	**83.4±10.7**	89.1±5.0
	U-Net [62]	31.1M	93.3±6.0	82.8±10.2	81.0±8.2	82.3±11.4	88.9±5.1
	ResEncL [35, 37]	102.0M	93.4±6.0	81.4±11.1	80.5±8.8	82.9±10.8	88.8±5.1
	ResEncL*	102.0M	93.5±5.9	88.0±7.3	80.5±8.7	82.8±11.1	89.5±7.8
Vision-Language	U-Net & CLIP [46]	19.1M	92.4±6.8	77.1±12.7	78.5±9.6	80.8±11.5	87.2±5.0
	Swin UNETR & CLIP [46]	62.2M	92.2±8.3	78.1±12.6	76.8±11.0	80.2±12.5	86.7±6.3
MONAI	LHU-Net [65]	8.6M	93.0±6.1	79.5±11.2	79.4±9.3	81.0±11.3	88.1±5.2
	UCTransNet [72]	68.0M	81.9±12.9	**86.5±8.0**	68.1±15.8	59.0±21.6	81.2±8.6
	Swin UNETR [68]	72.8M	90.5±8.6	77.2±15.1	75.4±11.8	75.6±14.5	84.9±7.1
	UNesT [85]	87.2M	90.9±7.3	77.7±16.1	74.4±11.8	76.2±12.1	85.0±6.2
	UNETR [25]	101.8M	88.8±8.4	76.5±16.4	71.5±12.8	72.3±14.5	83.4±7.0
	SegVol† [18]	181.0M	92.5±7.0	80.2±11.3	77.8±9.7	79.1±12.4	87.2±5.6
n/a	SAM-Adapter† [23]	11.6M	88.0±9.3	62.8±12.2	48.0±14.2	50.2±12.6	73.8±6.3
	MedFormer [19]	38.5M	93.4±6.4	82.1±11.7	80.7±10.1	83.1±11.2	89.0±5.4
	Diff-UNet [81]	434.0M	93.1±6.5	81.2±11.3	80.8±8.9	81.9±11.4	88.6±5.5

† These architectures were pre-trained (Appendix B.3).
* These architectures were trained on AbdomenAtlas 1.0 with enhanced label quality for the aorta and kidney classes (discussed in §4).

Results

Table 3: **Validation on TotalSegmentator (N=743).** Performances given as DSC score (mean±s.d.). For each class, we bold the best-performing results and highlight the runners-up, which show no significant difference from the best results at $p = 0.05$ level, in red. To ease the direct comparison with other literature, we also reported the *official* test set performance in Appendix Tables 11–12.

framework	architecture	param	spleen	kidneyR	kidneyL	gallbladder	liver
nnU-Net	UniSeg† [83]	31.0M	89.4±19.4	84.5±23.8	81.9±27.9	74.6±27.3	91.7±16.5
	MedNeXt [64]	61.8M	91.6±18.2	85.5±24.7	86.0±23.8	75.8±28.4	93.0±15.8
	NexToU [66]	81.9M	83.0±29.5	78.2±32.7	78.7±30.8	72.0±31.1	87.6±23.0
	STU-Net-B [34]	58.3M	92.3±15.3	87.1±20.2	86.8±22.1	**78.5±24.9**	93.0±13.9
	STU-Net-L [34]	440.3M	91.6±17.8	88.2±18.5	86.3±22.9	78.1±24.6	**94.2±11.2**
	STU-Net-H [34]	1457.3M	**92.4±14.6**	**88.9±16.2**	86.5±23.4	77.7±25.3	94.0±11.4
	U-Net [62]	31.1M	91.2±17.8	88.4±18.3	87.7±20.8	78.3±25.5	93.4±13.8
	ResEncL [35, 37]	102.0M	91.8±17.5	88.9±18.0	**88.2±20.5**	78.0±25.1	91.7±18.4
	ResEncL*	102.0M	92.0±16.7	89.9±15.3	89.5±18.3	78.0±24.7	92.4±17.4
Vision-Language	U-Net & CLIP [46]	19.1M	87.4±23.8	83.6±25.5	82.7±26.6	73.1±29.0	91.6±14.8
	Swin UNETR & CLIP [46]	62.2M	87.1±22.4	81.1±28.9	77.0±32.3	70.3±30.9	91.6±16.0
MONAI	LHU-Net [65]	8.6M	86.0±25.7	81.8±29.3	82.4±26.9	71.3±32.0	87.7±22.9
	UCTransNet [72]	68.0M	76.4±34.5	74.3±35.1	62.0±41.4	69.6±31.8	82.6±28.1
	Swin UNETR [68]	72.8M	66.3±36.4	59.7±39.3	58.5±40.1	50.6±40.5	80.2±28.7
	UNesT [85]	87.2M	79.5±26.6	73.8±32.3	72.0±33.8	50.3±39.9	87.6±20.8
	UNETR [25]	101.8M	60.4±37.9	47.9±39.5	41.9±39.7	40.0±36.7	78.1±29.8
	SegVol† [18]	181.0M	87.1±23.0	82.8±23.4	82.6±24.8	68.1±29.2	89.4±20.4
n/a	SAM-Adapter† [23]	11.6M	53.5±33.3	8.5±11.1	19.9±22.0	11.5±17.5	66.4±35.4
	MedFormer [19]	38.5M	90.7±15.0	85.5±18.4	84.0±21.5	74.1±26.7	92.8±12.4
	Diff-UNet [81]	434.0M	88.3±23.5	81.3±27.9	81.0±28.3	71.8±29.9	92.4±14.8

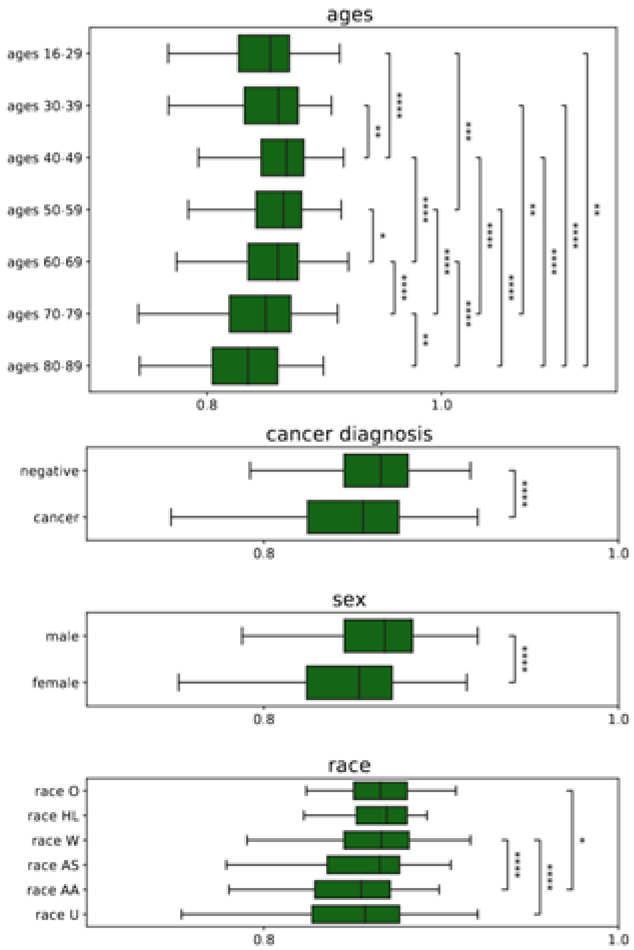
framework	architecture	param	stomach	aorta	IVC‡	pancreas	average
nnU-Net	UniSeg† [83]	31.0M	74.0±29.5	69.2±31.5	72.8±25.8	70.3±30.9	71.8±28.0
	MedNeXt [64]	61.8M	77.2±28.7	71.9±30.1	75.2±23.5	71.6±31.4	73.9±27.3
	NexToU [66]	81.9M	69.0±34.7	61.5±33.0	59.4±32.7	66.8±31.9	61.4±31.8
	STU-Net-B [34]	58.3M	78.6±26.5	74.2±28.9	77.3±19.5	74.9±27.4	76.6±24.9
	STU-Net-L [34]	440.3M	79.7±24.6	**75.7±26.9**	**77.6±18.7**	75.2±27.0	**78.9±21.5**
	STU-Net-H [34]	1457.3M	78.5±25.5	74.7±28.0	76.9±19.0	74.5±27.5	77.6±23.8
	U-Net [62]	31.1M	78.9±26.3	71.0±28.4	76.4±21.8	75.2±26.9	74.4±26.1
	ResEncL [35, 37]	102.0M	78.9±25.3	73.8±25.9	76.4±20.1	**76.3±25.8**	77.8±21.8
	ResEncL*	102.0M	80.9±23.0	84.2±20.5	76.3±20.0	77.3±24.9	84.5±20.1
Vision-Language	U-Net & CLIP [46]	19.1M	77.7±26.7	59.0±32.8	65.8±27.2	74.6±25.7	67.7±28.4
	Swin UNETR & CLIP [46]	62.2M	71.2±30.6	58.6±34.5	63.6±27.3	70.3±28.8	64.6±30.7
MONAI	LHU-Net [65]	8.6M	71.3±31.8	63.0±34.0	67.5±28.5	68.6±32.5	65.6±31.8
	UCTransNet [72]	68.0M	61.6±36.1	49.7±34.8	49.3±36.4	59.0±35.1	48.5±34.4
	Swin UNETR [68]	72.8M	52.2±35.1	54.5±36.9	38.1±34.6	42.3±34.4	45.4±31.1
	UNesT [85]	87.2M	63.9±31.4	54.7±36.9	38.9±36.2	50.0±32.9	49.4±32.3
	UNETR [25]	101.8M	42.1±32.0	41.0±31.3	41.3±32.3	28.2±29.1	37.3±27.9
	SegVol† [18]	181.0M	71.6±29.8	60.8±29.8	63.0±24.3	66.3±28.0	66.8±26.2
n/a	SAM-Adapter† [23]	11.6M	48.4±30.9	15.2±18.6	4.8±8.1	30.9±21.7	23.1±19.7
	MedFormer [19]	38.5M	**80.4±23.6**	70.3±28.0	70.0±24.4	72.5±27.9	75.1±24.1
	Diff-UNet [81]	434.0M	73.4±29.7	61.0±34.5	60.7±33.3	69.7±29.7	62.5±31.8

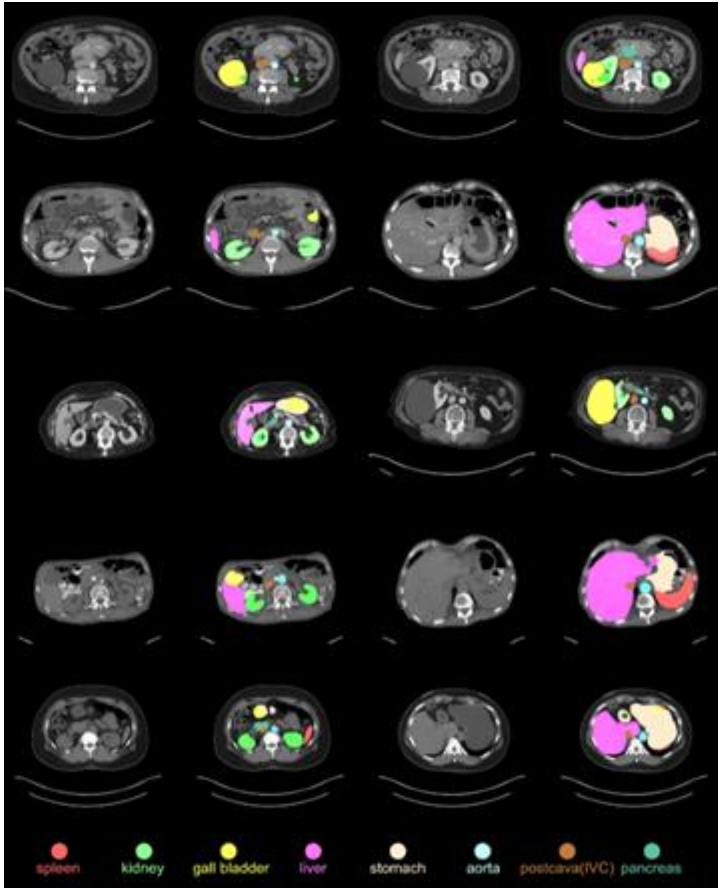
† These architectures were pre-trained (Appendix B.3).

‡ The class IVC (inferior vena cava) shares the same meaning as the class postcava in other datasets (e.g., AbdomenAtlas 1.0 and JHH).

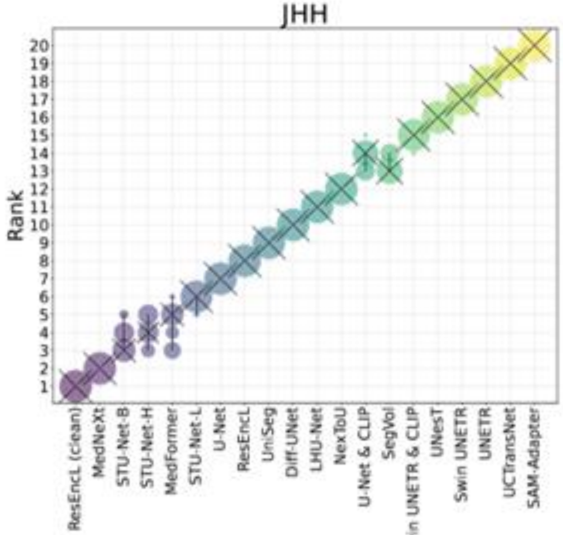
* These architectures were trained on AbdomenAtlas 1.0 with enhanced label quality for the aorta and kidney classes (discussed in §4).

Potential Confounders Significantly Impact AI

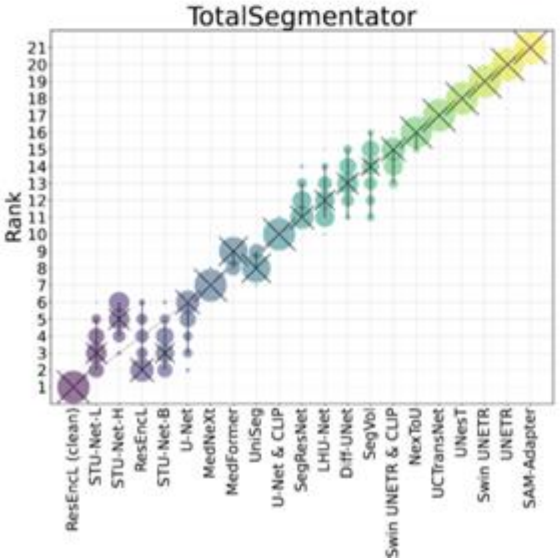




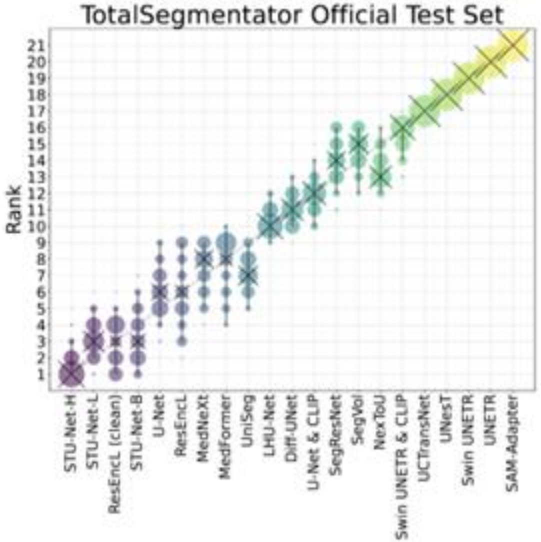
Test Set Size is Key



N = 5,160



N = 743



N = 59

Conclusions

1. OOD evaluation: AI performance varies significantly across OOD datasets
2. Large test datasets: more meaningful rankings and nuanced analysis
3. Per-organ analysis revealed AI strengths obscured by mean results
4. Per-group analysis revealed AI biases
5. With creator invitation and third-party evaluation, we establish a fair reference point for future AI algorithms

Touchstone 2.0 is accepting submissions, now with 9K+ CTs

Participate in Touchstone 2.0!

RSNA® 2024

Building Intelligent Connections

Technical Exhibits: Dec. 1–4

Thank You!