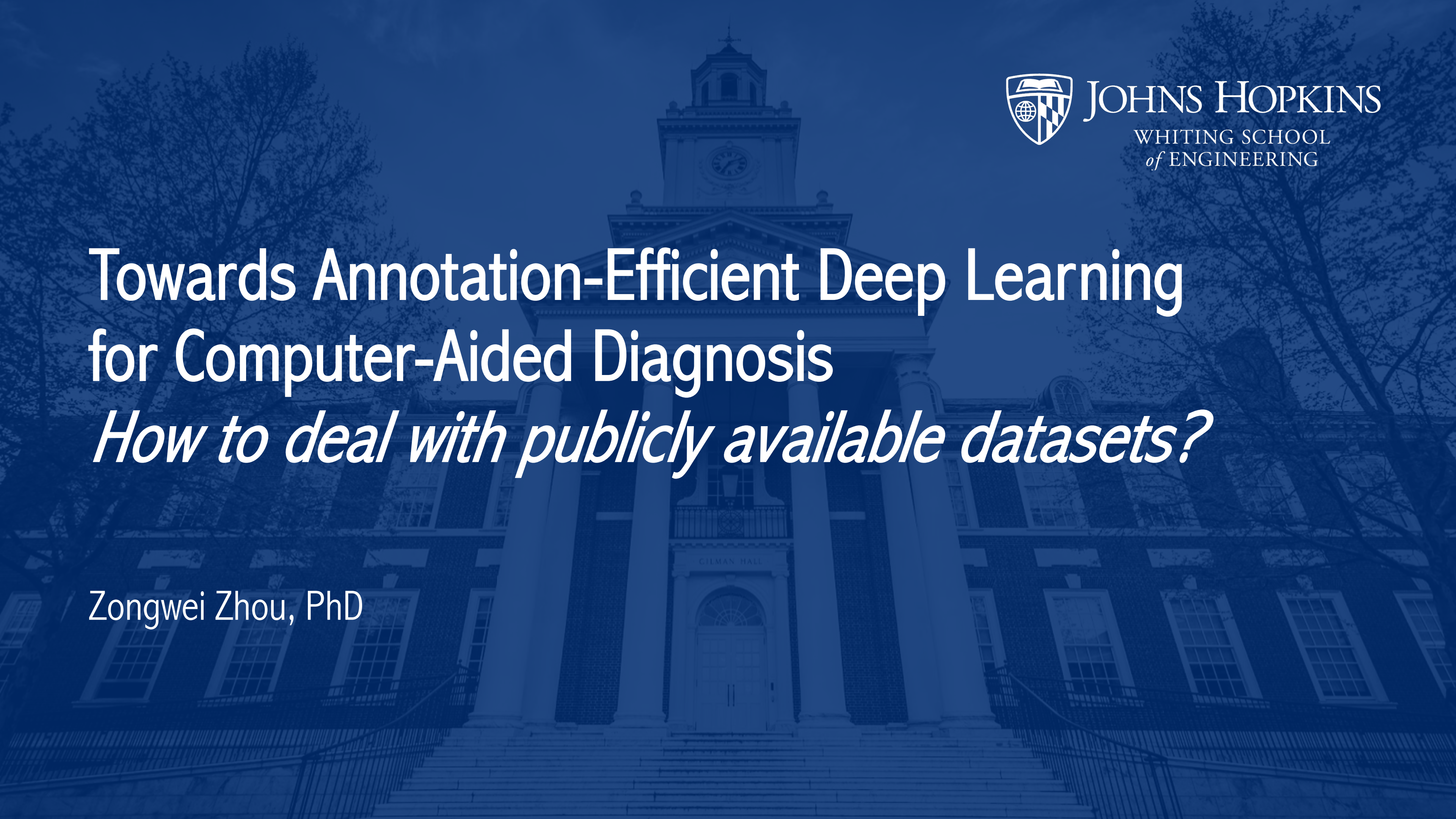


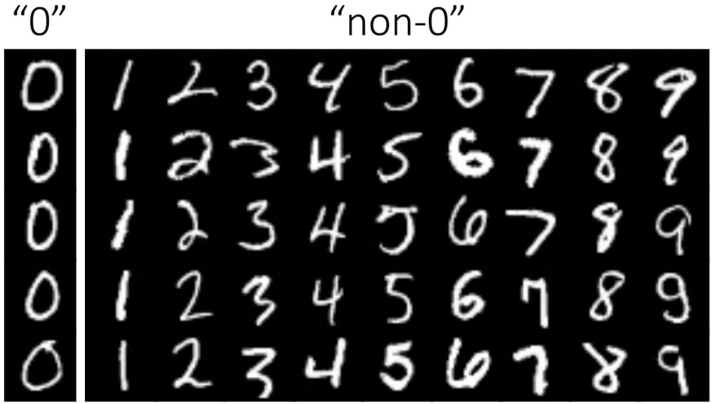
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis

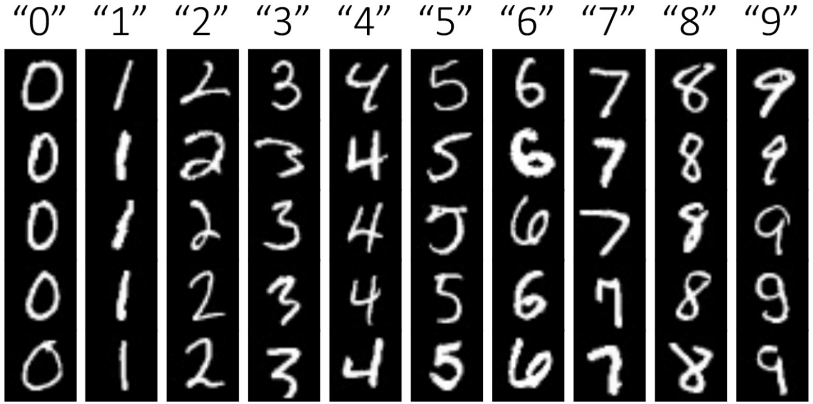
How to deal with publicly available datasets?

Zongwei Zhou, PhD

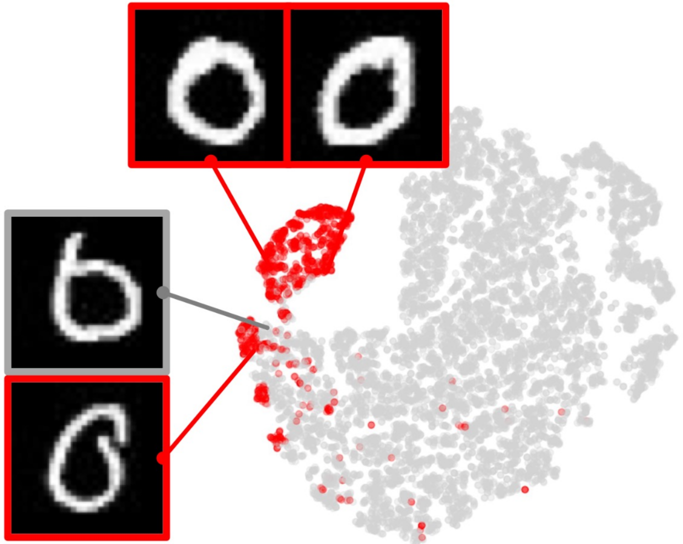
- Task: To classify images of the digit zero
- MNIST: A dataset that provides images and annotations of 0~9
- MNIST-*zero*: Derived from MNIST, wherein only the images of the digit zero are labeled as positives and the remainder are negatives (*sufficient for the task*)
- The total numbers of images are the same in MNIST and MNIST-*zero*
- *Which dataset would you prefer for the task, MNIST or MNIST-zero?*



MNIST-*zero*

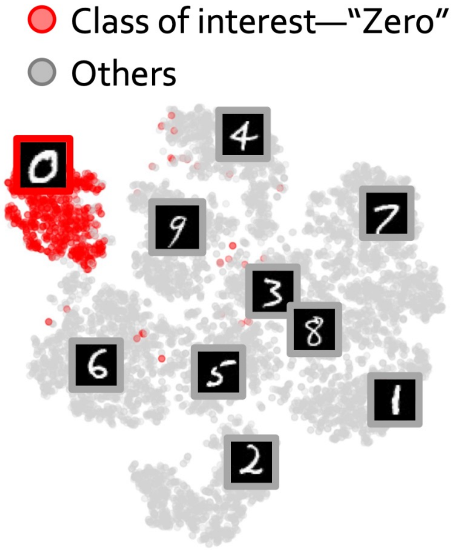


MNIST



MNIST-*zero*

AUC of “Zero”: 98.8%

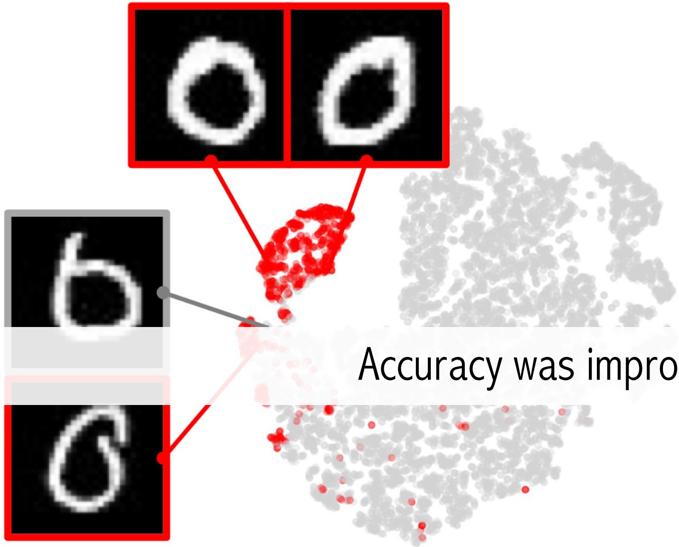


MNIST

AUC of “Zero”: 99.7%

1. Zhu, Z., Kang, M., Yuille, A. and Zhou, Z., Assembling Existing Labels from Public Datasets to Diagnose Novel Diseases: COVID-19 in Late 2019. Medical Imaging Meets NeurIPS 2022. https://www.cs.jhu.edu/~alanlab/Pubs22/zhu2022assembling.pdf
2. Kang, M., Lu, Y., Yuille, A.L. and Zhou, Z., 2021. Data, Assemble: Leveraging Multiple Datasets with Heterogeneous and Partial Labels. arXiv preprint arXiv:2109.12265.

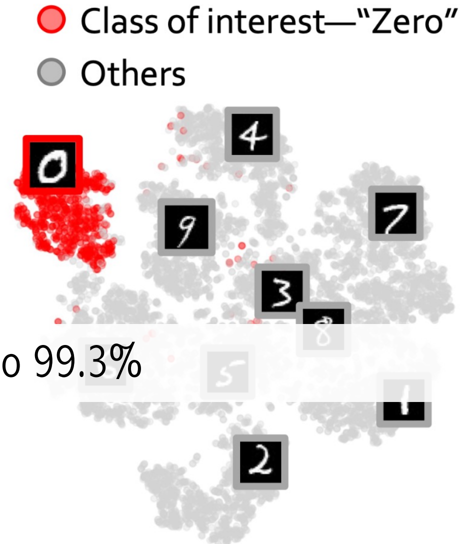
Accuracy was improved from 96.3% to 99.3%



MNIST-*zero*

AUC of “Zero”: 98.8%

COVID-19



MNIST

AUC of “Zero”: 99.7%

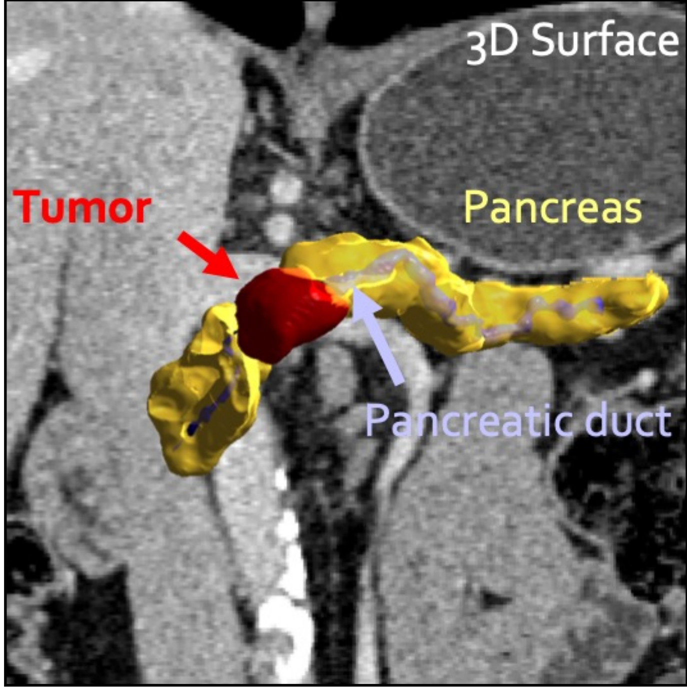
COVID-19 &

14 chest diseases from NIH ChestXray (2017)

1. Zhu, Z., Kang, M., Yuille, A. and Zhou, Z., Assembling Existing Labels from Public Datasets to Diagnose Novel Diseases: COVID-19 in Late 2019. Medical Imaging Meets NeurIPS 2022. https://www.cs.jhu.edu/~alanlab/Pubs22/zhu2022assembling.pdf
2. Kang, M., Lu, Y., Yuille, A.L. and Zhou, Z., 2021. Data, Assemble: Leveraging Multiple Datasets with Heterogeneous and Partial Labels. arXiv preprint arXiv:2109.12265.

Goal: Detecting and Segmenting Cancer

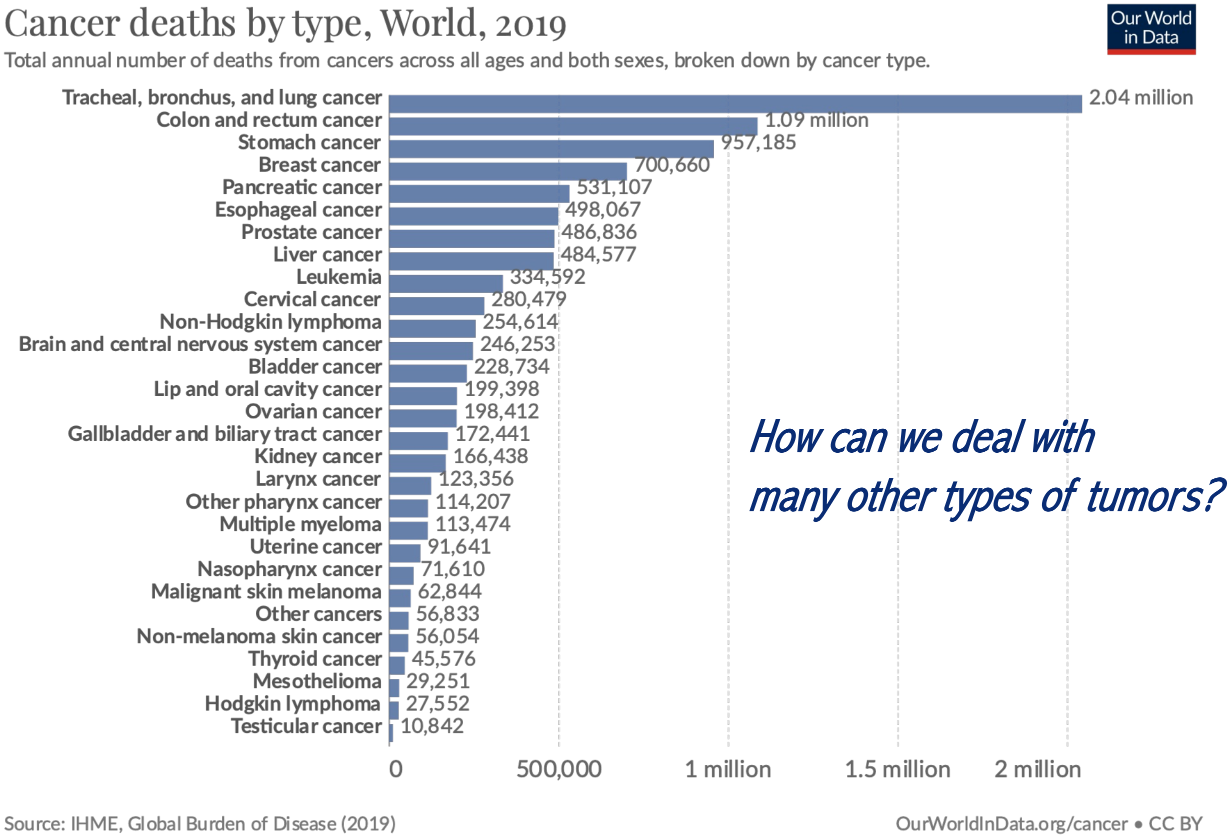
- **Detailed per-voxel annotations are limited in public datasets**
  - Colon tumors: 126 examples
  - Liver tumors: 131 examples
  - Pancreas tumors: 282 examples
  - Kidney tumors: 300 examples



- **High-performance AI algorithms require large annotated data**
  - Pancreas tumors: 5,038 annotated CT scans in FELIX ☛ Sensitivity=97%, Specificity=99%
  - This annotation took 15 human-year to create

[1] Xia, Y., Yu, Q., Chu, L., ... & Fishman, E. K. (2022). The FELIX Project: Deep Networks To Detect Pancreatic Neoplasms. medRxiv.

Goal: Detecting and Segmenting Cancers (Not Cancer)



Cancer deaths by type, World, 2019

Total annual number of deaths from cancers across all ages and both sexes, broken down by cancer type.

Source: IHME, Global Burden of Disease (2019)

OurWorldInData.org/cancer • CC BY

How can we deal with many other types of tumors?

Goal: Detecting and Segmenting Cancer<u>s</u> (Not Cancer)

- *How can we deal with many other types of tumors?*

- Two perspectives
- I. Exploiting existing public datasets and their partial annotation
- II. Exploring the potential of ultra-weak annotation (e.g., radiology report and synthetic tumors)

I will present our major achievements of the projects

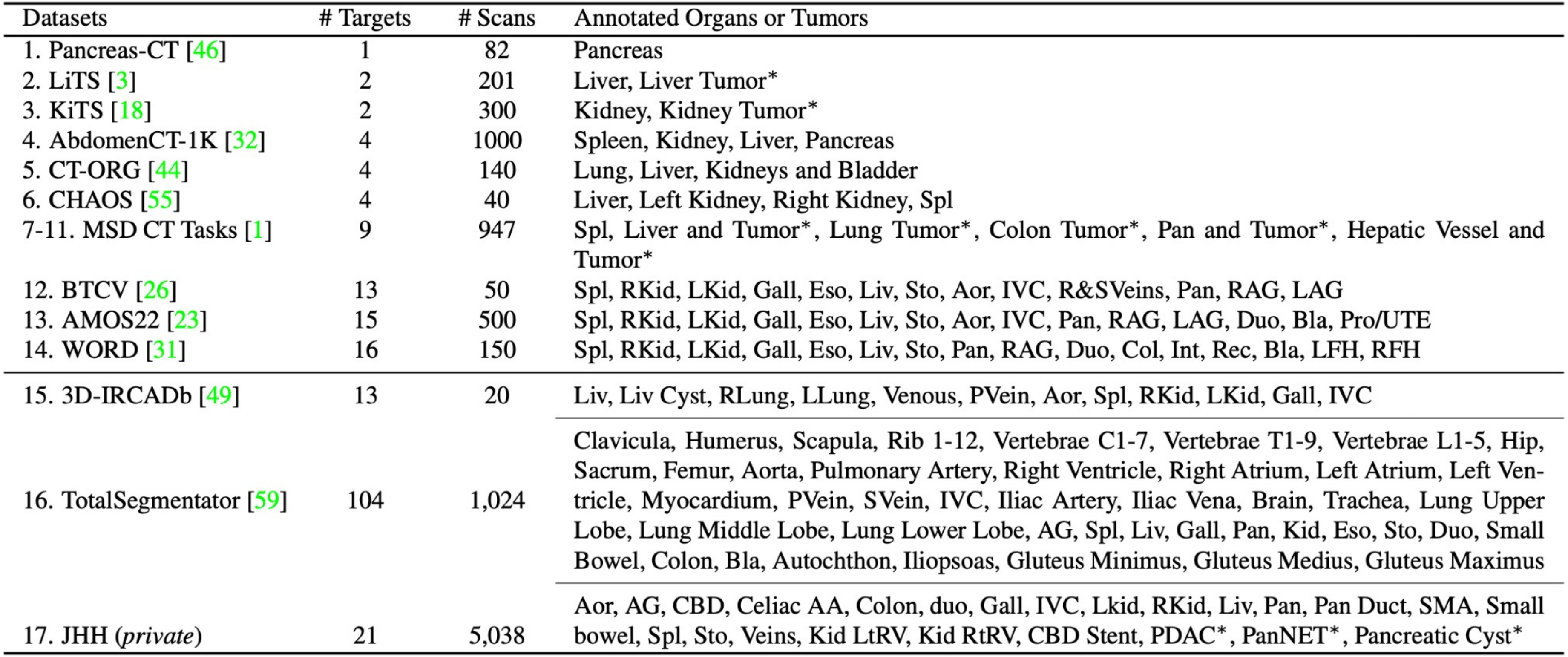
CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

Jie Liu[1], Yucheng Tang[2], Yixiao Zhang[3], Jie-Neng Chen[3], Junfei Xiao[3], Yongyi Lu[3], Yixuan Yuan[1], Alan Yuille[3], and Zongwei Zhou[3,*]

[1]City University of Hong Kong [2]NVIDIA [3]Johns Hopkins University

The first-place solution in Medical Segmentation Decathlon (MSD)

Publicly available abdominal CTs: *16 U-Nets* ☹



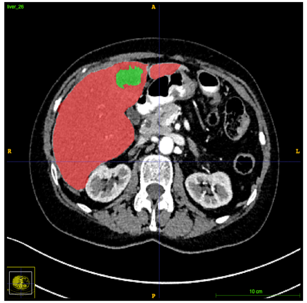
Datasets	# Targets	# Scans	Annotated Organs or Tumors
1. Pancreas-CT [46]	1	82	Pancreas
2. LiTS [3]	2	201	Liver, Liver Tumor*
3. KiTS [18]	2	300	Kidney, Kidney Tumor*
4. AbdomenCT-1K [32]	4	1000	Spleen, Kidney, Liver, Pancreas
5. CT-ORG [44]	4	140	Lung, Liver, Kidneys and Bladder
6. CHAOS [55]	4	40	Liver, Left Kidney, Right Kidney, Spl
7-11. MSD CT Tasks [1]	9	947	Spl, Liver and Tumor*, Lung Tumor*, Colon Tumor*, Pan and Tumor*, Hepatic Vessel and Tumor*
12. BTCV [26]	13	50	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&SVeins, Pan, RAG, LAG
13. AMOS22 [23]	15	500	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, Pro/UTE
14. WORD [31]	16	150	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH
15. 3D-IRCADb [49]	13	20	Liv, Liv Cyst, RLung, LLung, Venous, PVein, Aor, Spl, RKid, LKid, Gall, IVC
16. TotalSegmentator [59]	104	1,024	Clavicula, Humerus, Scapula, Rib 1-12, Vertebrae C1-7, Vertebrae T1-9, Vertebrae L1-5, Hip, Sacrum, Femur, Aorta, Pulmonary Artery, Right Ventricle, Right Atrium, Left Atrium, Left Ventricle, Myocardium, PVein, SVein, IVC, Iliac Artery, Iliac Vena, Brain, Trachea, Lung Upper Lobe, Lung Middle Lobe, Lung Lower Lobe, AG, Spl, Liv, Gall, Pan, Kid, Eso, Sto, Duo, Small Bowel, Colon, Bla, Autochthon, Iliopsoas, Gluteus Minimus, Gluteus Medius, Gluteus Maximus
17. JHH (*private*)	21	5,038	Aor, AG, CBD, Celiac AA, Colon, duo, Gall, IVC, Lkid, RKid, Liv, Pan, Pan Duct, SMA, Small bowel, Spl, Sto, Veins, Kid LtRV, Kid RtRV, CBD Stent, PDAC*, PanNET*, Pancreatic Cyst*

Goal: Segment everything in the abdomen

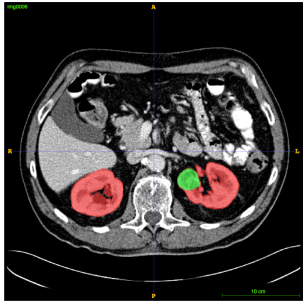
- **Approach:** Developing a single (Universal) model to learn from an assembly of public datasets
  - 2,995 CT scans; 25 organs; 6 tumors; 252 GB in total
- **Challenge I:** Domain gap between datasets
- **Challenge II:** Inconsistent annotation protocol and partial annotation
- **Challenge III:** Adapt to other organs/tumors

Illustration

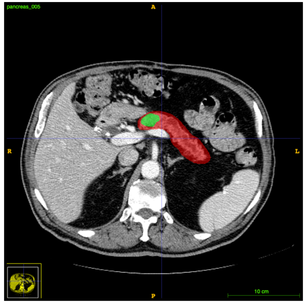
To segment major organs and to detect possible abnormalities



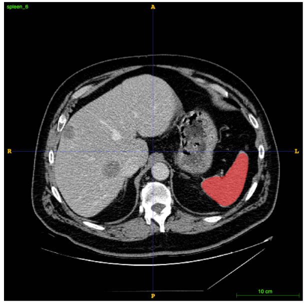
LiTS



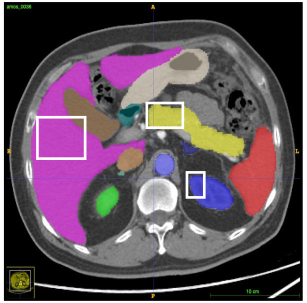
KiTS



MSD-Pancreas

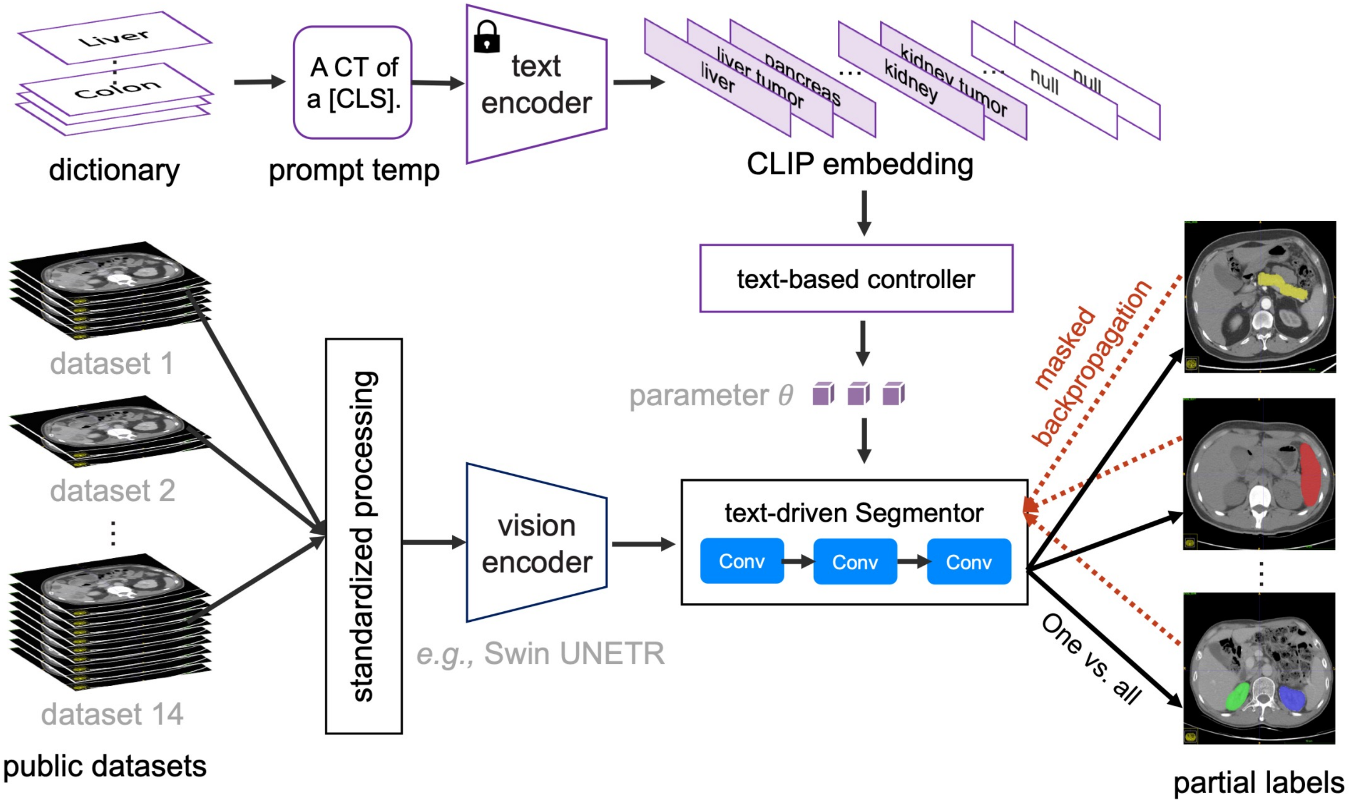


MSD-Spleen

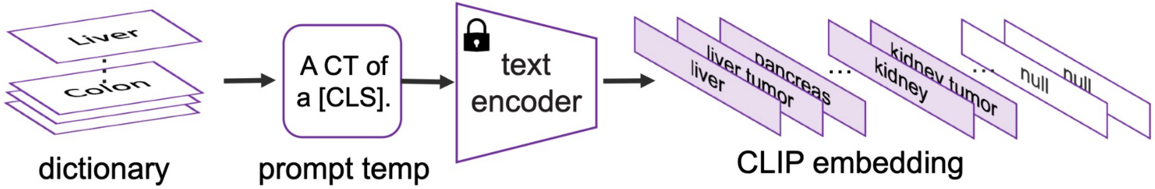


Ideal

The Universal Model



The Universal Model—why CLIP embedding?


Liver

Colon

dictionary

A CT of a [CLS].

prompt temp

text encoder

liver

liver tumor

pancreas

kidney

kidney tumor

null

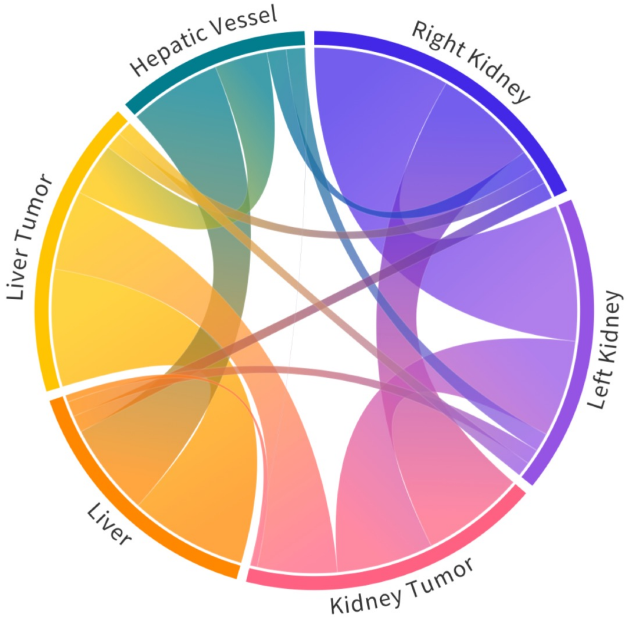
null

CLIP embedding

Conventional one-hot embedding

1. *No semantic meaning*
2. *Not extendable to novel classes*

liver:	[1,0,0,0,0,0]
liver tumor:	[0,1,0,0,0,0]
left kidney:	[0,0,1,0,0,0]
right kidney:	[0,0,0,1,0,0]
kidney tumor:	[0,0,0,0,1,0]
hepatic vessel:	[0,0,0,0,0,1]

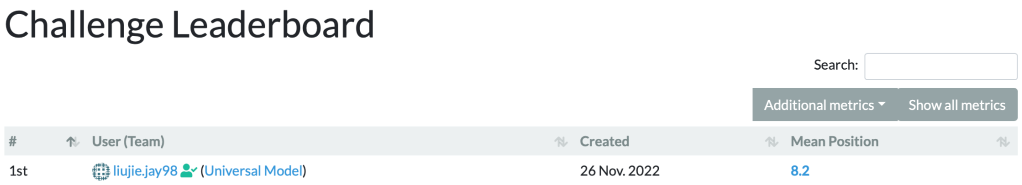


CLIP embedding

1. *Semantic meaning*
2. *Fixed length*

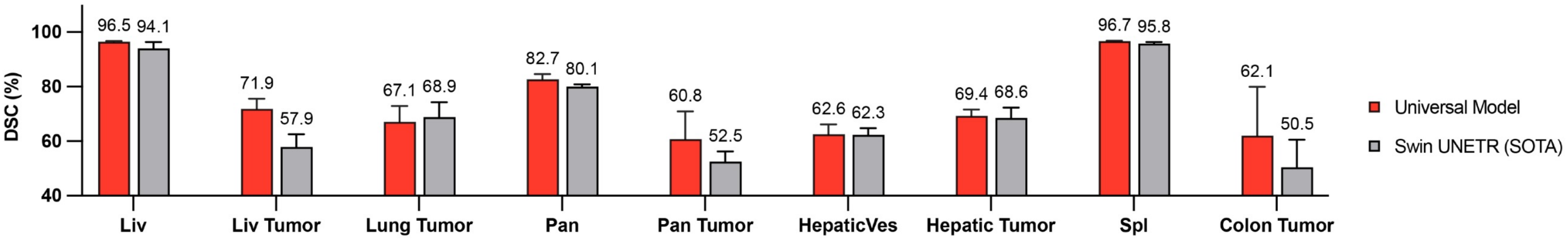
A1. Rank first in public datasets

- A performance demonstration on Medical Segmentation Decathlon (measured by DSC score)
- The improvement over the previous SOTA is quite significant



Challenge Leaderboard

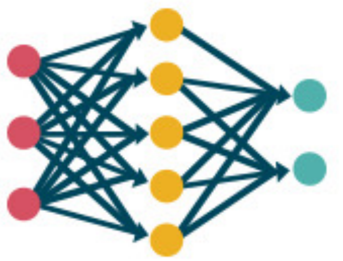
#	User (Team)	Created	Mean Position
1st	liujie.jay98 (Universal Model)	26 Nov. 2022	8.2



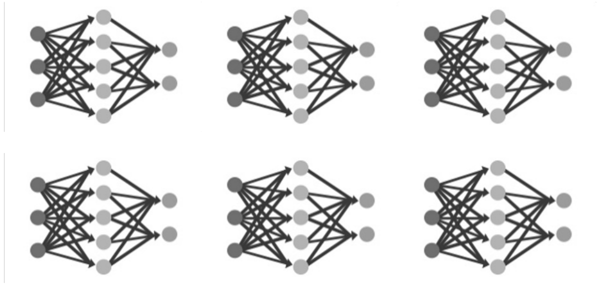
1. Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., ... & Hatamizadeh, A. (2022). Self-supervised pre-training of swin transformers for 3d medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20730-20740).

A2. Computation efficiency

- Universal Model is computationally efficient compared with dataset-specific models.



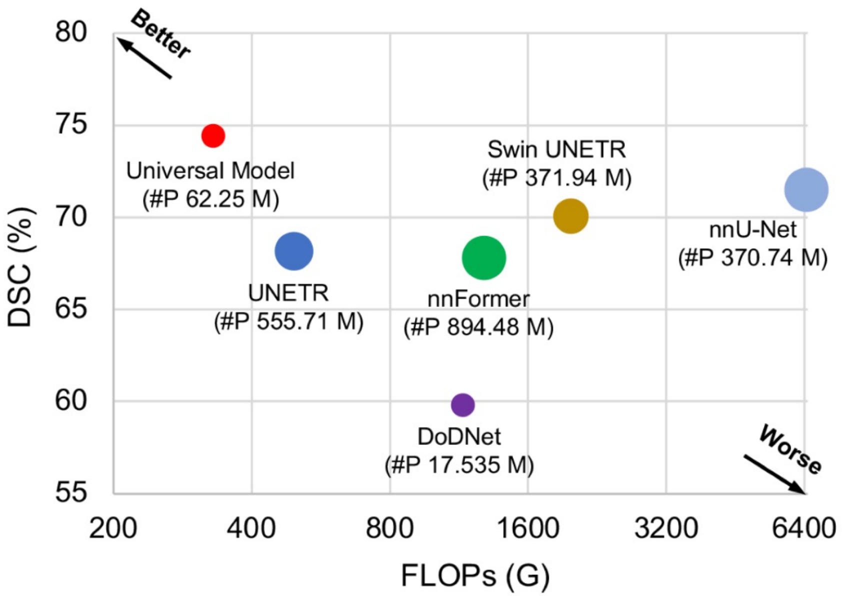
Ours: One for All (Ave: 14.22 s/scan)
Include load data time and inference time



Others: 6 models for 6 tasks (Ave: 190.26 s/scans)

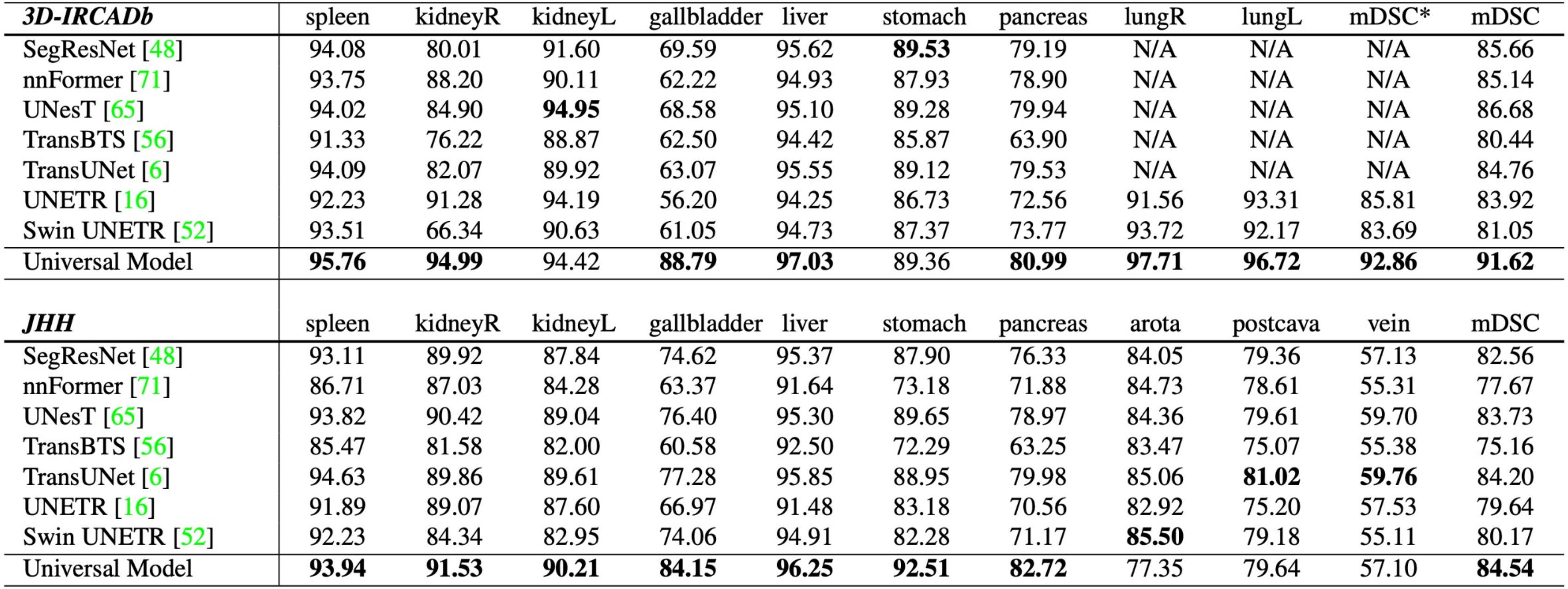
A2. Computation efficiency

- Universal Model is computationally efficient compared with dataset-specific models.
- 19x faster than nnU-Net (2nd best in performance) and 6x faster than Swin UNETR (3rd best)



A3. Generalize to other datasets

- Universal Model outperforms other dataset-specific models without being trained on those datasets.

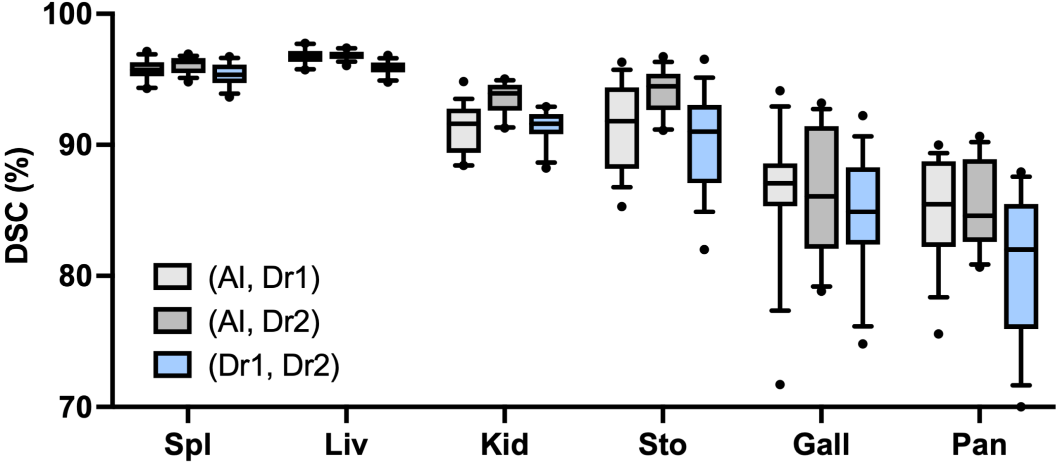


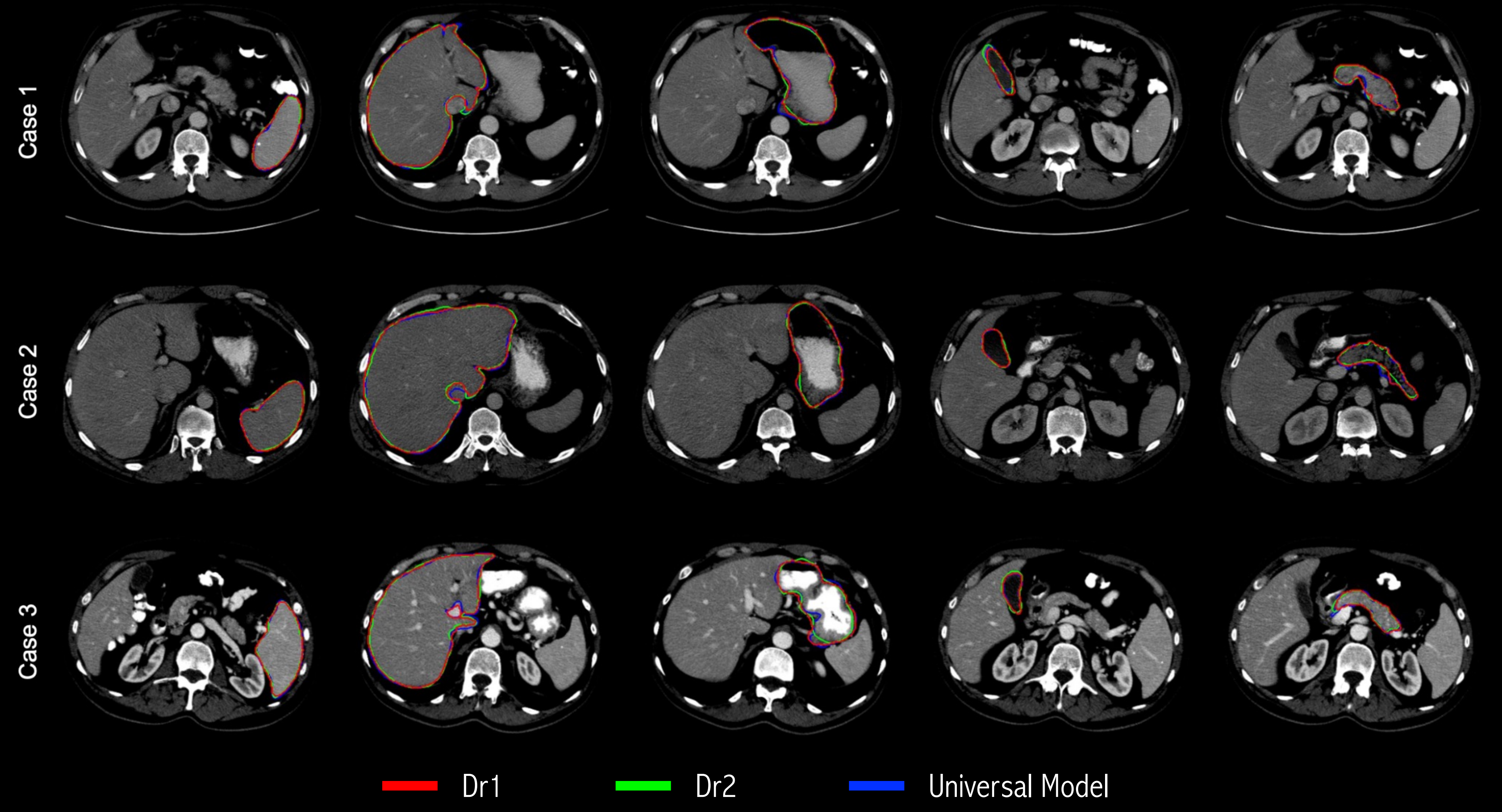
3D-IRCADb	spleen	kidneyR	kidneyL	gallbladder	liver	stomach	pancreas	lungR	lungL	mDSC*	mDSC
SegResNet [48]	94.08	80.01	91.60	69.59	95.62	**89.53**	79.19	N/A	N/A	N/A	85.66
nnFormer [71]	93.75	88.20	90.11	62.22	94.93	87.93	78.90	N/A	N/A	N/A	85.14
UNesT [65]	94.02	84.90	**94.95**	68.58	95.10	89.28	79.94	N/A	N/A	N/A	86.68
TransBTS [56]	91.33	76.22	88.87	62.50	94.42	85.87	63.90	N/A	N/A	N/A	80.44
TransUNet [6]	94.09	82.07	89.92	63.07	95.55	89.12	79.53	N/A	N/A	N/A	84.76
UNETR [16]	92.23	91.28	94.19	56.20	94.25	86.73	72.56	91.56	93.31	85.81	83.92
Swin UNETR [52]	93.51	66.34	90.63	61.05	94.73	87.37	73.77	93.72	92.17	83.69	81.05
Universal Model	**95.76**	**94.99**	94.42	**88.79**	**97.03**	89.36	**80.99**	**97.71**	**96.72**	**92.86**	**91.62**

JHH	spleen	kidneyR	kidneyL	gallbladder	liver	stomach	pancreas	arota	postcava	vein	mDSC
SegResNet [48]	93.11	89.92	87.84	74.62	95.37	87.90	76.33	84.05	79.36	57.13	82.56
nnFormer [71]	86.71	87.03	84.28	63.37	91.64	73.18	71.88	84.73	78.61	55.31	77.67
UNesT [65]	93.82	90.42	89.04	76.40	95.30	89.65	78.97	84.36	79.61	59.70	83.73
TransBTS [56]	85.47	81.58	82.00	60.58	92.50	72.29	63.25	83.47	75.07	55.38	75.16
TransUNet [6]	94.63	89.86	89.61	77.28	95.85	88.95	79.98	85.06	**81.02**	**59.76**	84.20
UNETR [16]	91.89	89.07	87.60	66.97	91.48	83.18	70.56	82.92	75.20	57.53	79.64
Swin UNETR [52]	92.23	84.34	82.95	74.06	94.91	82.28	71.17	**85.50**	79.18	55.11	80.17
Universal Model	**93.94**	**91.53**	**90.21**	**84.15**	**96.25**	**92.51**	**82.72**	77.35	79.64	57.10	**84.54**

A4. High-quality pseudo labels

- We demonstrate the pseudo label quality (AI) for the six organs is comparable to human annotators (Dr1, Dr2)
  - *If we spend a lot more money to ask radiologists to annotate these six organs, it might turn out that our pseudo labels can do a similar quality annotation (which is a waste of money and time).*



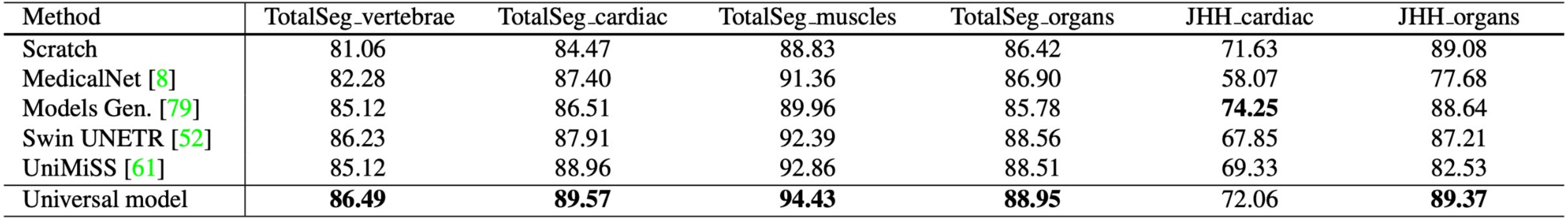

Case 1
Case 2
Case 3
🟥 Dr1
🟩 Dr2
🟦 Universal Model

A4. High-quality pseudo labels

- We demonstrate the pseudo label quality (AI) for the six organs is comparable to human annotators (Dr1, Dr2)
  - *If we spend a lot more money to ask radiologists to annotate these six organs, it might turn out that our pseudo labels can do a similar quality annotation (which is a waste of money and time).*
- We have completed the missing labels in 14 public datasets and will release a dataset of 3,410 CT scans with six organs annotated by high-quality pseudo labels. (Some refinement of pseudo labels is required)
  - *We encourage the research community to concentrate on creating datasets of the harder organs/tumors*

A5. Transferability to downstream tasks

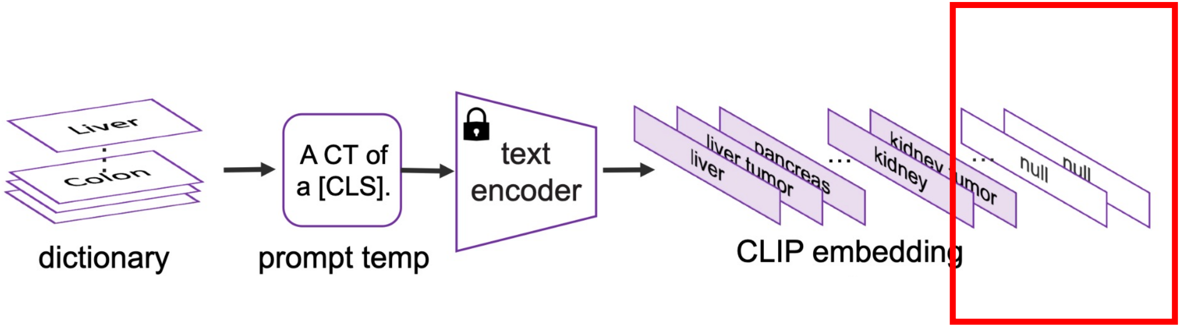
- Universal Model can be used for fine-tuning, performing better than many famous medical Foundation Models
- The benefit of existing self-supervised learning for downstream tasks is indirect
  - *New scheme: pre-training by segmenting*



Method	TotalSeg_vertebrae	TotalSeg_cardiac	TotalSeg_muscles	TotalSeg_organs	JHH_cardiac	JHH_organs
Scratch	81.06	84.47	88.83	86.42	71.63	89.08
MedicalNet [8]	82.28	87.40	91.36	86.90	58.07	77.68
Models Gen. [79]	85.12	86.51	89.96	85.78	**74.25**	88.64
Swin UNETR [52]	86.23	87.91	92.39	88.56	67.85	87.21
UniMiSS [61]	85.12	88.96	92.86	88.51	69.33	82.53
Universal model	**86.49**	**89.57**	**94.43**	**88.95**	72.06	**89.37**

Looking forward

- Participate in upcoming MICCAI, RSNA, Grand Challenges for medical image segmentation
  - Generalizability, transferability, computational efficiency
- Continual and incremental learning for novel classes that will be annotated in the future



e.g., other fine-grained types of cancer

Synthetic Tumors Make AI Segment Tumors Better

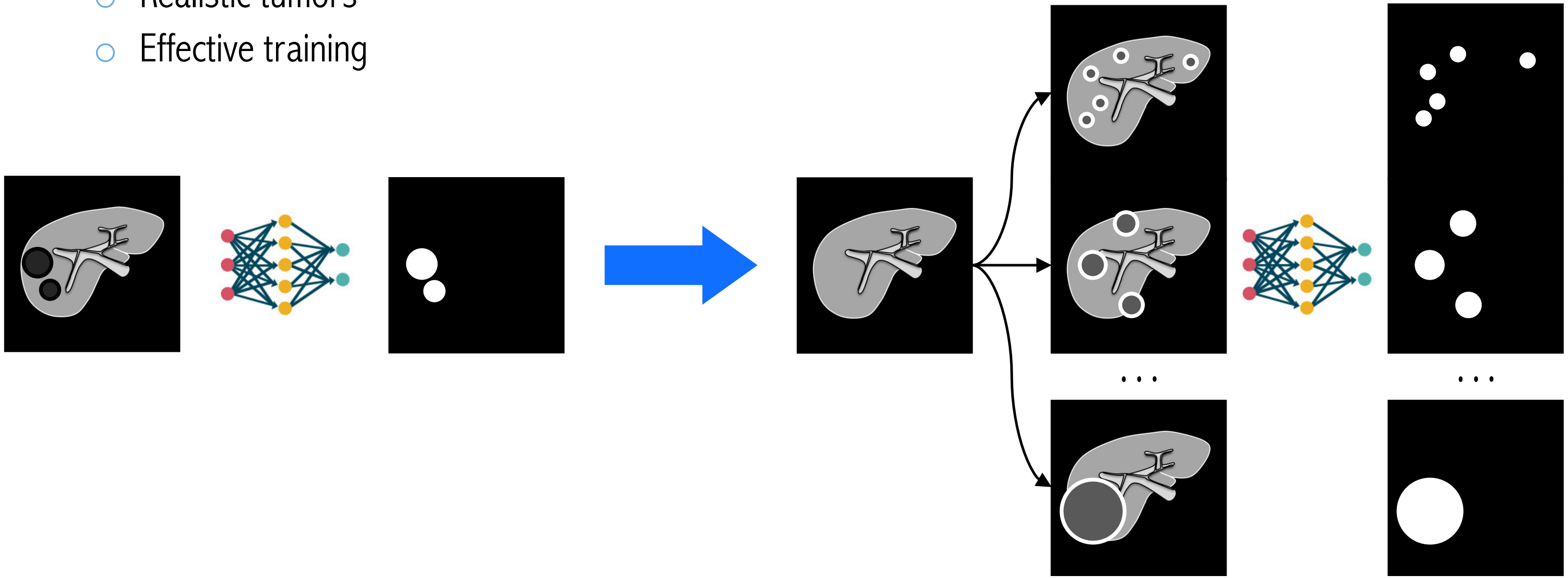
Qixin Hu[1], Yixiong Chen[2], Junfei Xiao[3], Shuwen Sun[4], Jie-Neng Chen[3],
Alan Yuille[3], and Zongwei Zhou[3,*]

[1]Huazhong University of Science and Technology [2]Fudan University
[3]Johns Hopkins University [4]The First Affiliated Hospital of Nanjing Medical University

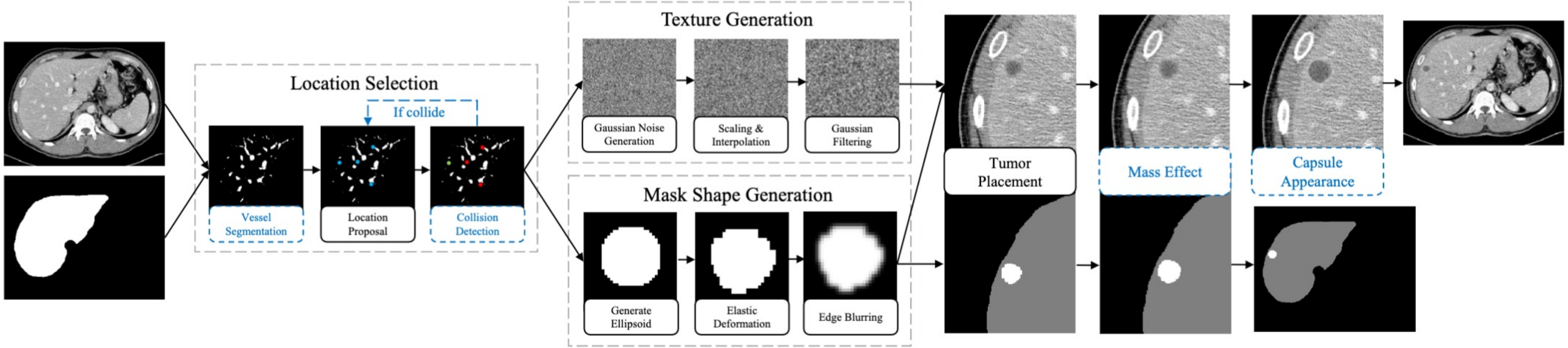
Github: https://github.com/MrGiovanni/SyntheticTumors

Training paradigm shift

- Old paradigm: AI models segment tumors from images (label-intensive)
- New paradigm: Tumors are generated for AI models to segment them (label-free)
  - Realistic tumors
  - Effective training



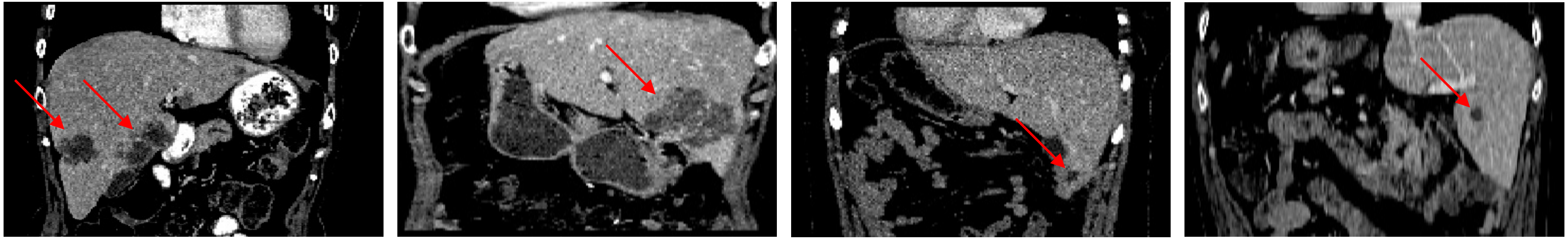
Liver tumor generator



1. Hu, Q., Xiao, J., Chen, Y., ... & Zhou, Z. (2022). “Synthetic Tumors Make AI Segment Tumors Better.” Medical Imaging Meets NeurIPS, 2022.

Liver tumor generator

- How to (empirically) make the synthetic tumors more realistic?
  - Position prior: >60% of the tumors are on the lowest 1/3 of the liver
  - Shape prior: larger tumors usually have more irregular shapes
  - Color and texture prior: larger tumors are usually brighter with richer textures

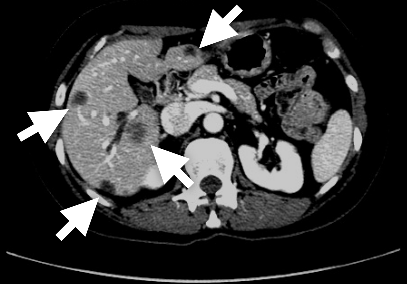


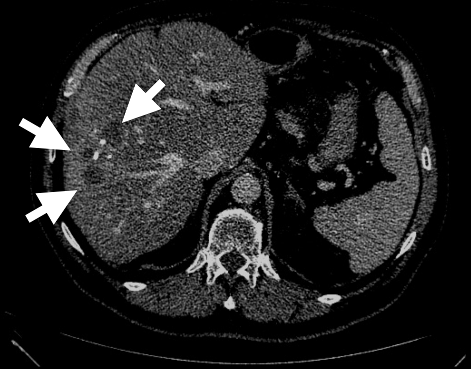
A1. Small domain gap between real and synthetic tumors

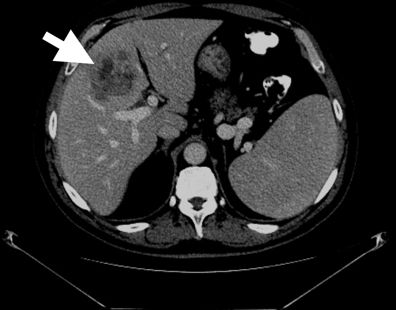
- We estimate the domain gap by two measures
- (I) Vision Turing Test[1]
  - Performed by two *medical professionals (6-year and 30-year experience)*
  - A total of 50 CT scans are used: 30 are real, 20 are synthetic (professionals do not know this)
  - Medical professionals must assign “real (1)”, “synthetic (-1)”, or “cannot tell (0)” to each CT scan

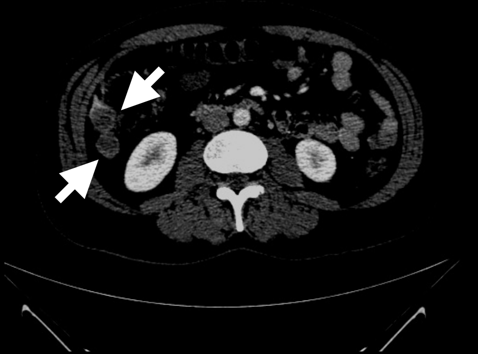
[1] Geman, D., Geman, S., Hallonquist, N. and Younes, L., 2015. Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences, 112(12), pp.3618-3623.

Medical professionals with over 6-year experience cannot tell which are real and which are synthetic tumor with an accuracy of 20% (*lower than random guess*)

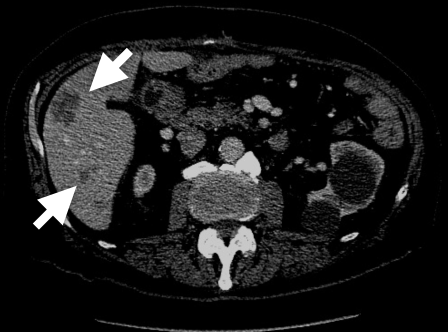


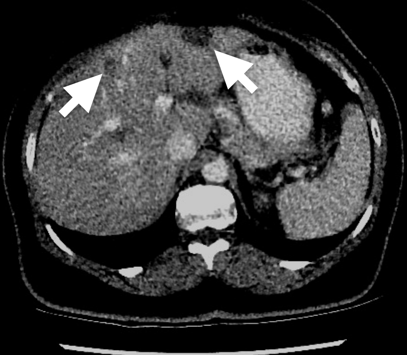


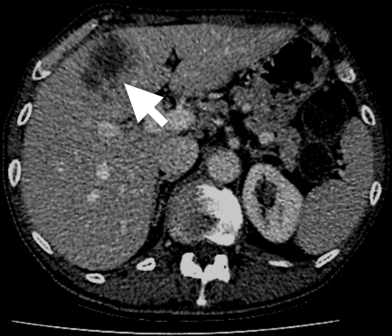


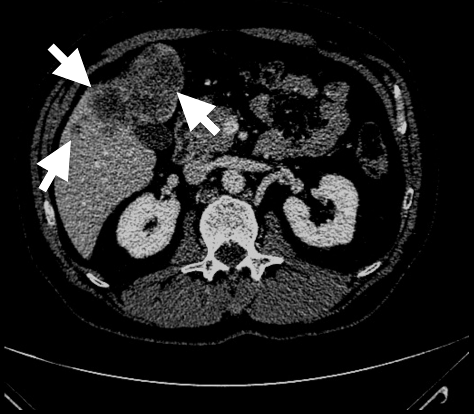


Can you?









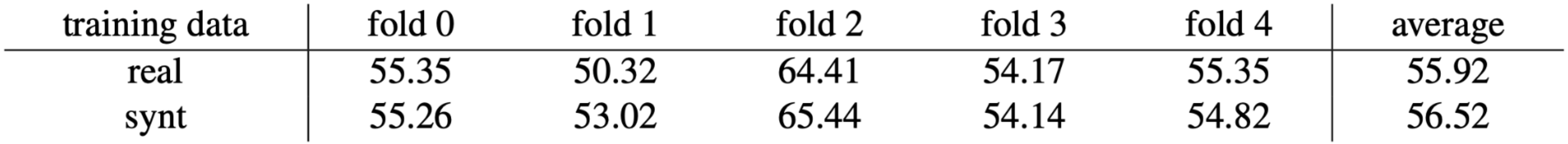
A1. Small domain gap between real and synthetic tumors

- We estimate the domain gap by two measures
- (II) Quantitative evaluation on the tests set of real and synthetic tumors.
  - Test on real tumors: 22 CT scans from LiTS
  - Test on synthetic tumors: 22 CT scans from CT-ORG

	Test on real tumors	Test on synthetic tumors
AI trained with *real* tumors	52.3	
AI trained with *synthetic* tumors	52.0	

A2. AI trained with synthetic tumors ≈ with real tumors

- The quantitative result is exciting because no previous synthetic tumor has achieved a similar or even close performance to real tumors.
- Essentially, we won the liver tumor segmentation challenge (MSD-Liver) while not using any annotation provided by this challenge, outperforming top teams who trained AI using 101 annotated CT scans.



training data	fold 0	fold 1	fold 2	fold 3	fold 4	average
real	55.35	50.32	64.41	54.17	55.35	55.92
synt	55.26	53.02	65.44	54.14	54.82	56.52

real: previous top 1 team on MSE challenge (Swin UNETR Base).

1. Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., ... & Hatamizadeh, A. (2022). Self-supervised pre-training of swin transformers for 3d medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20730-20740).

Training AI on synthetic tumors performs almost as well as training it on real tumors.

CT

AI prediction
trained on real tumors
with per-voxel annotation

AI prediction
trained on synthetic tumors
with no annotation

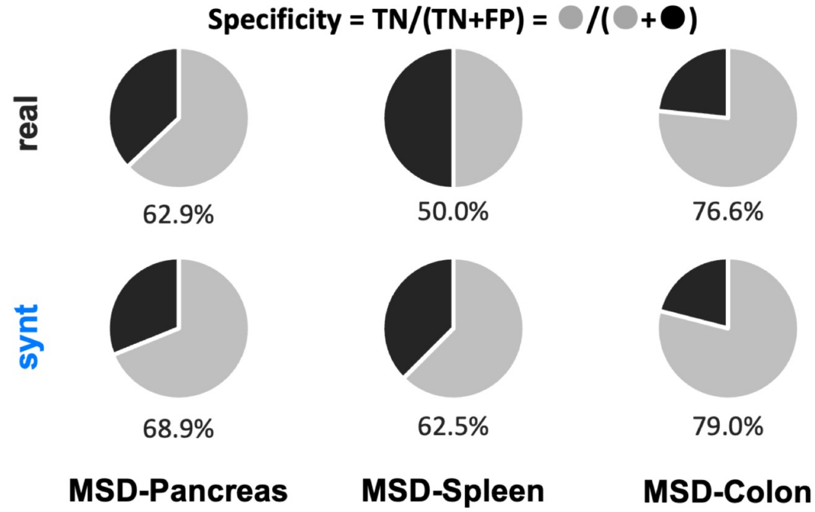
● Liver

● Liver tumor

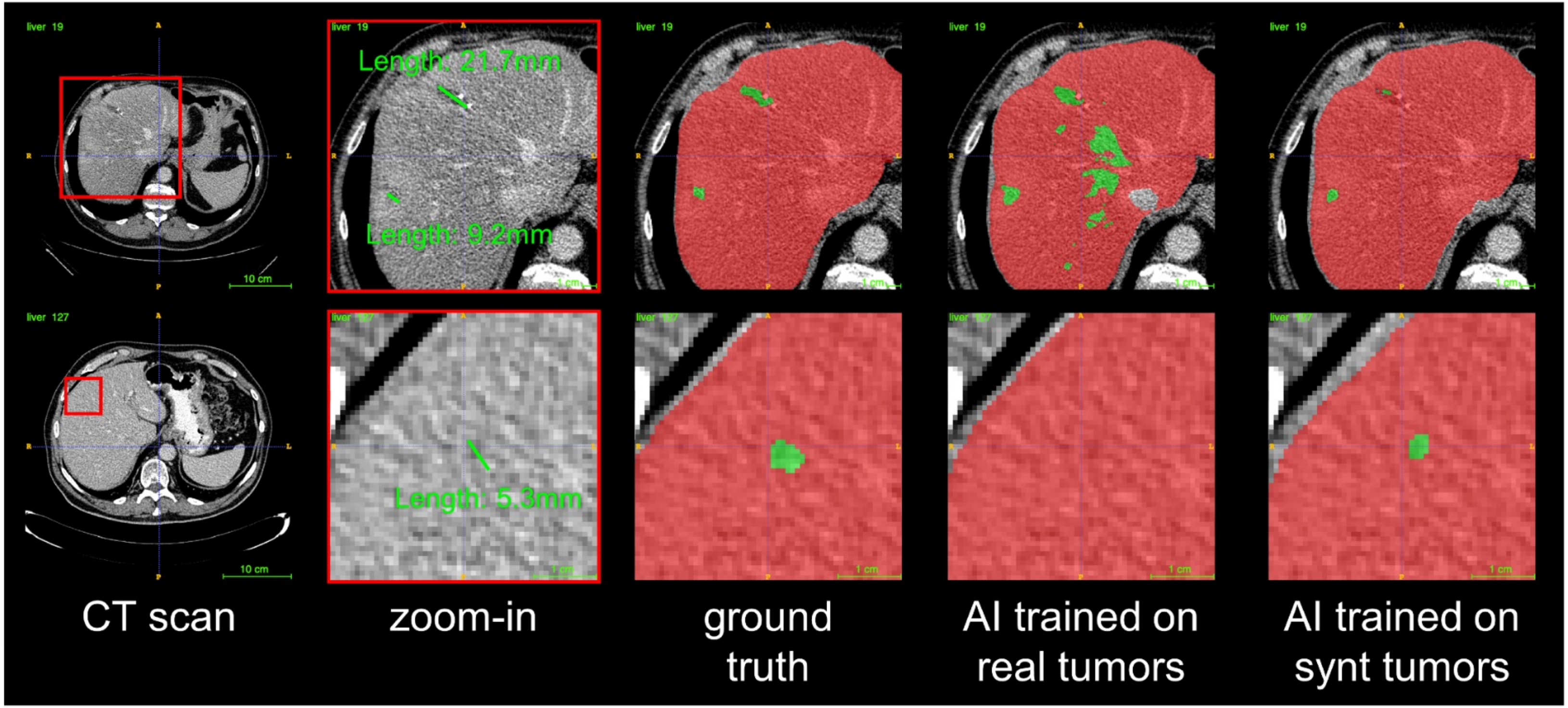
[1] Hu, Q., Xiao, J., Chen, Y., … & Zhou, Z. (2022). “Synthetic Tumors Make AI Segment Tumors Better.” Medical Imaging Meets NeurIPS, 2022.

A3. AI trained with synthetic tumors generates less FPs

- The tumor dataset usually provides a lot more positive examples than negative examples. Although the model is good at detecting liver tumors, it offers a low specificity on the healthy CT scans in the inference.
- Ours: The datasets are diverse, consisting of a large number positive and negative examples (as control).

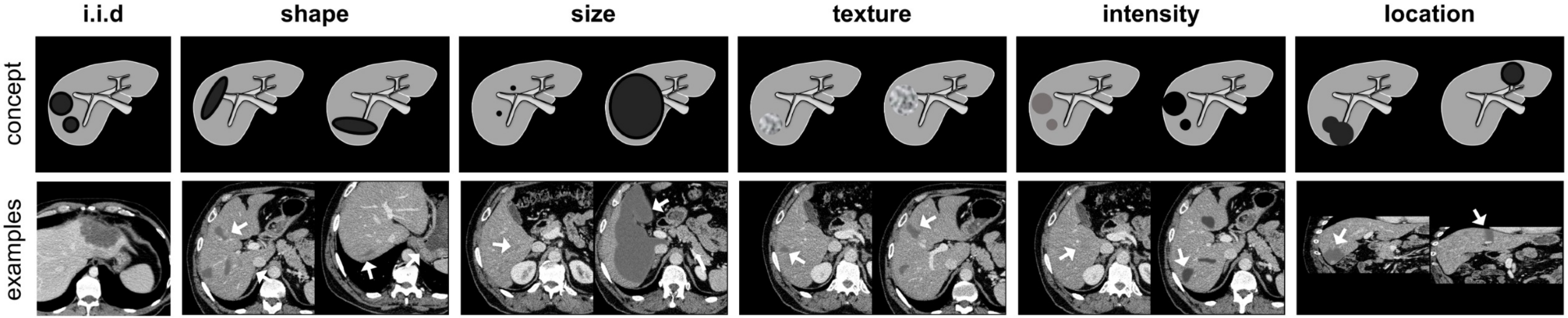


A4. AI trained with synthetic tumors can detect tiny tumors



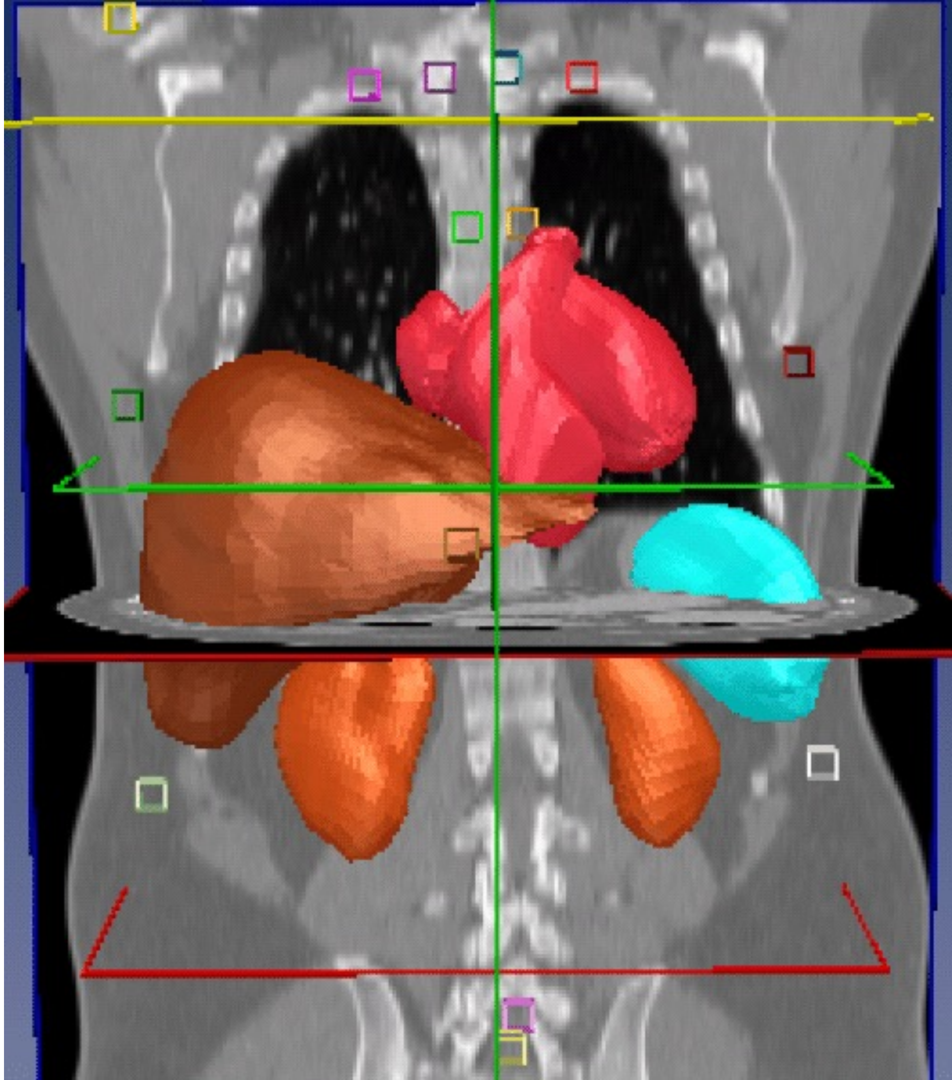
A5. Controllable robustness benchmark

- The limitations of AI models in tumor segmentation are not fully studied.
  - There are only 70 CT scans available for evaluating AI in MSD-Liver
- Synthetic tumors enable us to perform an extensive evaluation of these models in segmenting liver tumors that vary from different conditions.
  - Shape, size, texture, intensity, location, etc.



Looking forward

- We plan to generate synthetic tumors in many more organs
- In the future, annotations are still needed, but these annotations will be only used for evaluation
    - Colon tumors: 126 examples
    - Liver tumors: 131 examples
    - Pancreas tumors: 282 examples
    - Kidney tumors: 300 examples
    - More fine-grained tumor types…

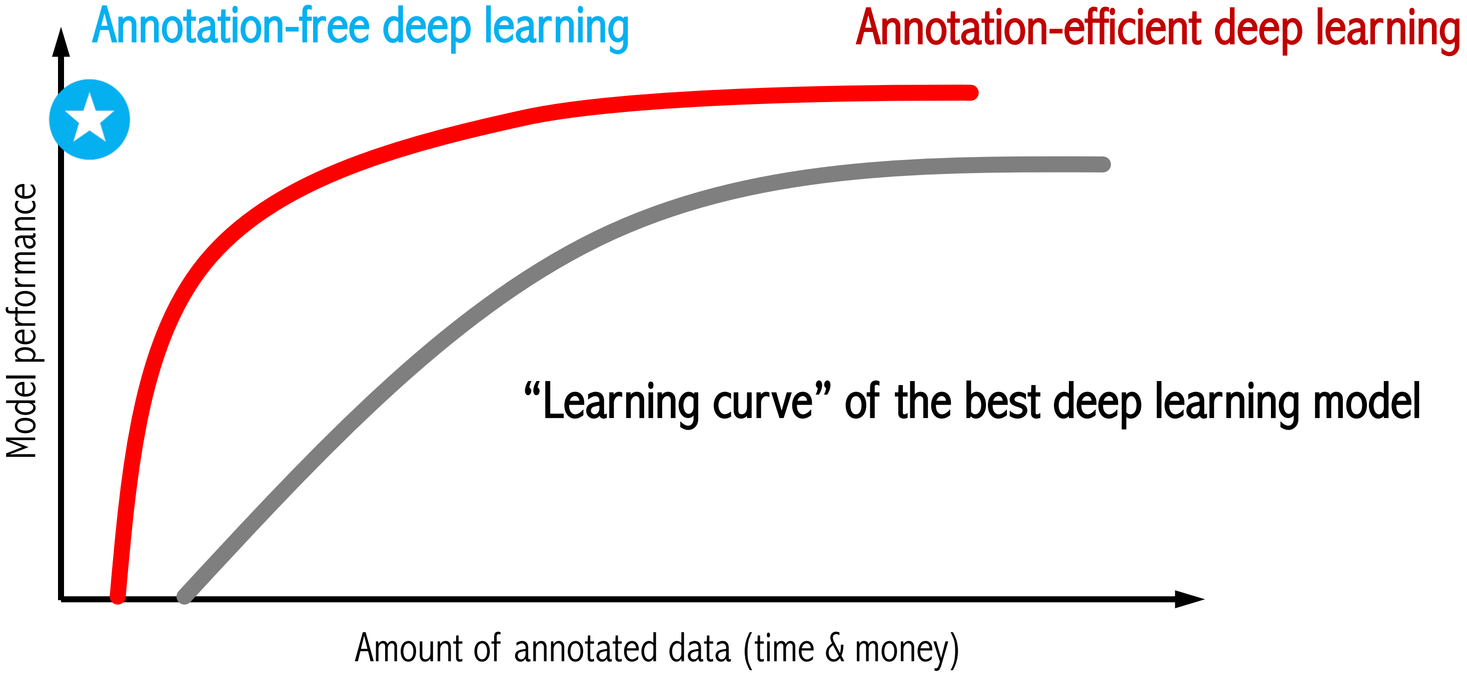


Summary

- Detecting and Segmenting Cancers (Not Cancer)
  - *How can we deal with many other types of tumors?*

- Two perspectives
- I. Exploiting existing public datasets and their partial annotation
  - *Universal Model GitHub: coming soon*
  - *Label-Assemble GitHub: https://github.com/MrGiovanni/LabelAssemble*
- II. Exploring the potential of ultra-weak annotation (e.g., synthetic tumors)
  - *Synthetic Tumors GitHub: https://github.com/MrGiovanni/SyntheticTumors*

Towards Annotation-Efficient *(-Free)* Deep Learning



1. Zhou, Z. (2021). Towards annotation-efficient deep learning for computer-aided diagnosis (Doctoral dissertation, Arizona State University).