Synthetic Tumors Make AI Segment Tumors Better

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

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

Code and Visual Turing Test: https://github.com/MrGiovanni/SyntheticTumors

Abstract

We develop a novel strategy to generate synthetic tumors. Unlike existing works, the tumors generated by our strategy have two intriguing advantages: (1) realistic in shape and texture, which even medical professionals can confuse with real tumors; (2) effective for AI model training, which can perform liver tumor segmentation similarly to a model trained on real tumors—this result is *unprecedented* because no existing work, using synthetic tumors only, has thus far reached a similar or even close performance to the model trained on real tumors. This result also implies that manual efforts for developing per-voxel annotation of tumors (which took years to create) can be considerably reduced for training AI models in the future. Moreover, our synthetic tumors have the potential to improve the success rate of small tumor detection by automatically generating enormous examples of small (or tiny) synthetic tumors.

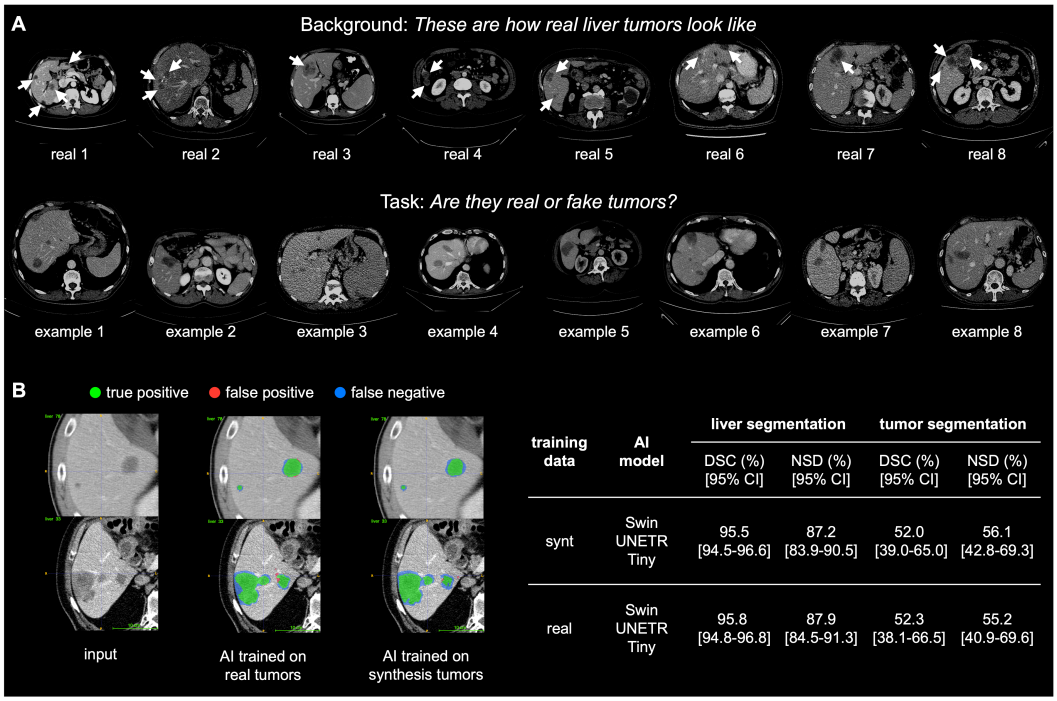
1 Introduction

Artificial intelligence (AI) has dominated medical image segmentation [21, 22, 7], but training an AI model (e.g., U-Net [13]) often requires a large number of detailed per-voxel annotations. Annotating medical images is not only expensive and time-consuming, but also requires extensive medical expertise, and sometimes needs the assistance of radiology reports and biopsy results to precisely annotate a tumor [20, 17]. Due to its high annotation cost, only a total of roughly 100 CT scans with annotated liver tumors are publicly available (provided by LiTS [1]) for training and testing models.

Generating synthetic tumors is an attractive research topic. There are some early attempts at generating COVID-19 infections on Chest CT scans [19], abdominal tumors in CT scans [8], diabetic lesions on retinal images [16], brain tumors on MRI images [18], and cancers in fluorescence microscopy images [6]. However, the synthetic tumors in those existing studies appear very different from the real tumors, and AI models trained using synthetic tumors perform significantly worse than those trained using real tumors due to the pronounced domain gap between real and synthetic tumors. *What makes synthesizing tumors so hard?* There are several important factors: shape, intensity, size, location, and most importantly, texture. In this paper, we develop a hand-crafted heuristic strategy to synthesize liver tumors in abdominal CT scans. Our synthetic tumors are realistic—even medical professionals can confuse them with real tumors in the Visual Turing Test [3, 4] (Figure 1A). Besides, AI models trained on our synthetic tumors can segment real tumors similar to those trained on real tumors with expensive, detailed per-voxel annotation. As shown in Figure 1B, the model trained on our (label-free) synthetic tumors achieves a Dice Similarity Coefficient (DSC) of 52.0% for segmenting real liver tumors, whereas AI trained on real tumors obtains a DSC of 52.3% (no statistical difference between the two performances). These results are unprecedented because no existing work, using synthetic tumors *only*, has thus far reached a similar performance to the model trained on real tumors.

*Corresponding author: Zongwei Zhou (zzhou82@jh.edu)

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



training data	AI model	liver segmentation		tumor segmentation	
		DSC (%) [95% CI]	NSD (%) [95% CI]	DSC (%) [95% CI]	NSD (%) [95% CI]
synt	Swin UNETR Tiny	95.5 [94.5-96.6]	87.2 [83.9-90.5]	52.0 [39.0-65.0]	56.1 [42.8-69.3]
real	Swin UNETR Tiny	95.8 [94.8-96.8]	87.9 [84.5-91.3]	52.3 [38.1-66.5]	55.2 [40.9-69.6]

Figure 1: **A.** We conduct an examination for humans to distinguish synthetic tumors from the real ones (i.e., Visual Turing Test). In these examples, some CT scans contain real liver tumors, and some contain synthetic tumors generated by our algorithm. **B.** Qualitative assessment and quantitative analysis of AI models trained on synthesis tumors and real tumors. Both of them can obtain similar segmentation performance while generating synthetic tumors requires no manual annotation cost.

More importantly, synthesizing tumors also enables us to exhaustively generate tumors of desired locations, sizes, shapes, textures, and intensity, which are not limited to a fixed finite-size training set. For example, it is hard to gather sufficient training examples with annotated small tumors since they often occur at the early stage of cancer, and present subtle abnormal textures to human eyes, and therefore it is difficult for experts to specify the boundary of the tumor manually. In contrast, our synthesizing strategy can generate enormous examples with small (or tiny) tumors, so it can potentially detect small tumors more effectively (see Figure 2). In summary, our ultimate goal is ambitious: to train AI models for tumor segmentation without using any manual annotation—this study makes a significant step towards it.

2 Materials & Methods

Dataset & Metric. Detailed per-voxel annotations of liver tumors are provided in the LiTS dataset [1]. The tumor types include HCC and secondary liver tumors and metastasis derived from colorectal, breast, and lung cancer. The volume of liver tumors ranges from 38mm^3 to 349 cm^3, and the radius of tumors is approximately in the range of [2, 44]mm. We split LiTS into a training set (101 CT scans) and a test set (22 CT scans), which follows the conventional setting used in the literature (e.g., [15]). An AI model (i.e., Swin UNETR-Tiny[2] [5]) is trained on the 101 CT scans with annotated liver tumors. For comparison, a dataset of 109 CT scans with a healthy liver is assembled from CHAOS [9] (20 CT scans), BTCV [10] (47 CT scans), and Pancreas-CT [14] (42 CT scans). We then generate liver tumors in these scans on the fly, resulting in enormous image-label pairs of synthetic tumors and their masks for training the AI model. To evaluate the model's segmentation performance, we calculate the Dice similarity coefficient (DSC) and Normalized Surface Dice (NSD) with 2mm tolerance to quantify the performance.

[2]Swin UNETR is a hybrid segmentation architecture, which integrates the benefits of both U-Net [13] and Transformer [2, 11]. We base our experiments on Swin UNETR because it is very competitive and has ranked top one in numerous public benchmarks [15], including liver tumor segmentation (MSD-Liver).

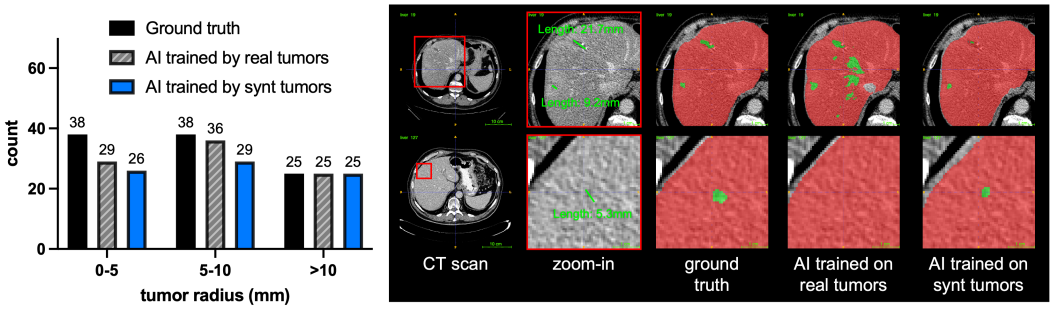


Figure 2: Performance of liver tumor detection stratified by tumor size. The left panel presents the tumor-level sensitivity. For both models, the false negatives are mostly smaller than 10mm. The right panel presents two examples of small tumors. The smallest tumor we detected was 2mm.

Tumor Generator. We develop a hand-crafted heuristic strategy to generate synthetic liver tumors, consisting of shape generation and texture generation. First, ellipse generation, elastic deformation, and mask blurring are sequentially applied to generate the sphere-like shape of tumors. Second, salt noise, Gaussian filtering, scaling, and clipping are applied to generate tumor-like texture. This is the basic version of our tumor generator, wherein the hyper-parameters are adjusted by (1) visual inspection between the real and synthetic tumors, (2) standardized guidance of the Liver Imaging Reporting and Data System (LI-RADS) [12], and (3) feedback from clinicians during the early iterations of the Visual Turing Test (introduced below). We have implemented several critical feedback into our tumor generator, e.g., the mass effect and cirrhosis on the healthy part of the liver around the synthetic tumor; the satellite effect of multiple small tumors around a large tumor.

Visual Turing Test is adopted to assess the quality of synthetic tumors from professionals' perspectives (examples in Figure 1A). The generated synthetic tumors are considered realistic if the professionals fail to distinguish them from the real tumors. Our preliminary results show an accuracy of 60% (30/50), performed by a professional with 6-year experience, but a more evaluation is required.

3 Results & Discussion

Both qualitative and quantitative results in Figure 1B demonstrate that the model trained on synthetic tumors performs similarly to the model trained on real tumors in segmenting real liver tumors from unseen CT scans. Specifically, the Swin UNETR-Tiny achieves DSC scores of 52.0% [95% CI: 39.0%-65.0%] and 52.3% [95% CI: 38.1%-66.5%] when trained on synthetic and real tumors, respectively. A slightly higher NSD score (56.1% vs. 55.2%) achieved by the model trained on synthetic tumors indicates that the model can also detect the boundary of the liver tumor precisely. Moreover, we evaluate the performance of small tumor detection. Figure 2 stratifies tumors by different sizes and plots the detection rates of models trained on synthetic and real tumors. Both models are capable of detecting liver tumors that are larger than 10mm radius. As of now, the model trained on real tumors (65 out of 76 real tumors detected) outperforms the model trained on synthetic tumors (55 out of 76 detected) in detecting tumors smaller than 10mm. We anticipate the ability of small tumor detection can be improved by generating more small-sized synthetic tumors in the training stage. This is one of the advantages of synthetic tumors because CT scans with small liver tumors are very difficult to collect and annotate in clinical practice. The right panel of Figure 2 presents two examples of small tumors that are successfully detected by the model, and visually, the segmentation quality is greater than the ground truth.

Conclusion. In this paper, we proposed an unsupervised strategy to generate realistic shapes and textures of liver tumors. Synthetic tumors enable AI models to perform similarly to the model trained on real liver tumors—collecting and annotating real tumors can take years to complete, but our proposed strategy is label-free. This reveals the great potential for to use of synthesis tumors to train AI models on larger-scale healthy CT datasets (which are much easier to obtain than CT scans with liver tumors). In practice, we can generate enormous synthetic tumors in CT scans, which facilitate annotation-efficient AI development and allow us to assess AI's capability of detecting tumors of varying locations, sizes, shapes, intensities, and textures.

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank Camille Torrico and Alexa Delaney for improving the writing of this paper.

References

[1] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015.

[4] Changhee Han, Yoshiro Kitamura, Akira Kudo, Akimichi Ichinose, Leonardo Rundo, Yujiro Furukawa, Kazuki Umemoto, Yuanzhong Li, and Hideki Nakayama. Synthesizing diverse lung nodules wherever massively: 3d multi-conditional gan-based ct image augmentation for object detection. In *2019 International Conference on 3D Vision (3DV)*, pages 729–737. IEEE, 2019.

[5] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022.

[6] Izabela Horvath, Johannes Paetzold, Oliver Schoppe, Rami Al-Maskari, Ivan Ezhov, Suprosanna Shit, Hongwei Li, Ali Ertürk, and Bjoern Menze. Metgan: Generative tumour inpainting and modality synthesis in light sheet microscopy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 227–237, 2022.

[7] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[8] Qiangguo Jin, Hui Cui, Changming Sun, Zhaopeng Meng, and Ran Su. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. *Knowledge-Based Systems*, 218:106753, 2021.

[9] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[10] B Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. 2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge, 2015. `doi:10.7303/syn3193805`.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[12] Guilherme M. Cunha, Kathryn J Fowler, Alexandra Roudenko, Bachir Taouli, Alice W Fung, Khaled M Elsayes, Robert M Marks, Irene Cruite, Natally Horvat, Victoria Chernyak, et al. How to use li-rads to report liver ct and mri observations. *RadioGraphics*, 41(5):1352–1367, 2021.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[14] Holger Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. Data from pancreas-ct, 2016. The Cancer Imaging Archive. `https://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU`.

[15] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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*, pages 20730–20740, 2022.

[16] Hualin Wang, Yuhong Zhou, Jiong Zhang, Jianqin Lei, Dongke Sun, Feng Xu, and Xiayu Xu. Anomaly segmentation in retinal images with poisson-blending data augmentation. *Medical Image Analysis*, 81:102534, 2022.

[17] Meiyun Wang, Fangfang Fu, Bingjie Zheng, Yan Bai, Qingxia Wu, Jianqiang Wu, Lin Sun, Qiuyu Liu, Mingge Liu, Yichen Yang, et al. Development of an ai system for accurately diagnose hepatocellular carcinoma from computed tomography imaging data. *British Journal of Cancer*, 125(8):1111–1121, 2021.

[18] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022.

[19] Qingsong Yao, Li Xiao, Peihang Liu, and S Kevin Zhou. Label-free segmentation of covid-19 lesions in lung ct. *IEEE transactions on medical imaging*, 40(10):2808–2819, 2021.

[20] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021.

[21] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[22] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.