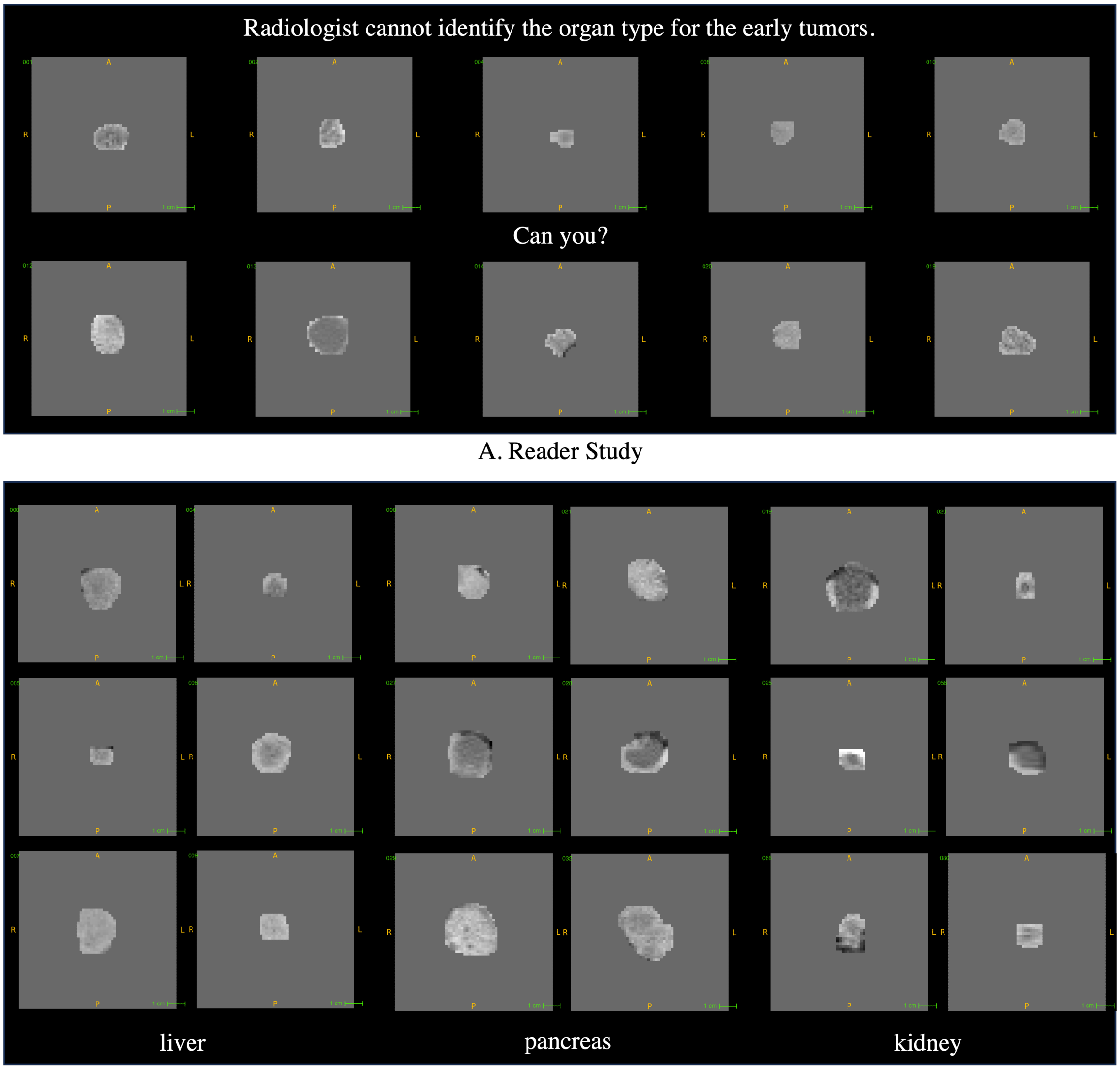
Towards Generalizable Tumor Synthesis
Supplementary Material

This supplementary material is organized as follows: §A provides visual examples for reader study and Visual Turing Test. §B provides a description of Radiomics Features. §C provides additional results for generalizable to multiple organs. §D provides additional results for generalizable to different patient demographics. §E provides the details of used datasets and implementation for DiffTumor and Segmentation Model. §F provides discussions about comparison with related works, unrealistic generation, and challenging case analysis.

A. Visual Examples



Radiologist cannot identify the organ type for the early tumors.
Can you?
liver
pancreas
kidney

A. Reader Study

B. More examples for reader study

Figure 8. **Visual examples for the reader study. A.** Quick test for identifying the organ type for the early tumors. **B.** More early tumors for three abdominal organs. The organs corresponding to the early tumor in the first row of **A** are the liver, pancreas, kidney, liver, and kidney. The organs corresponding to the early tumor in the second row of **A** are the pancreas, pancreas, liver, liver, and pancreas.

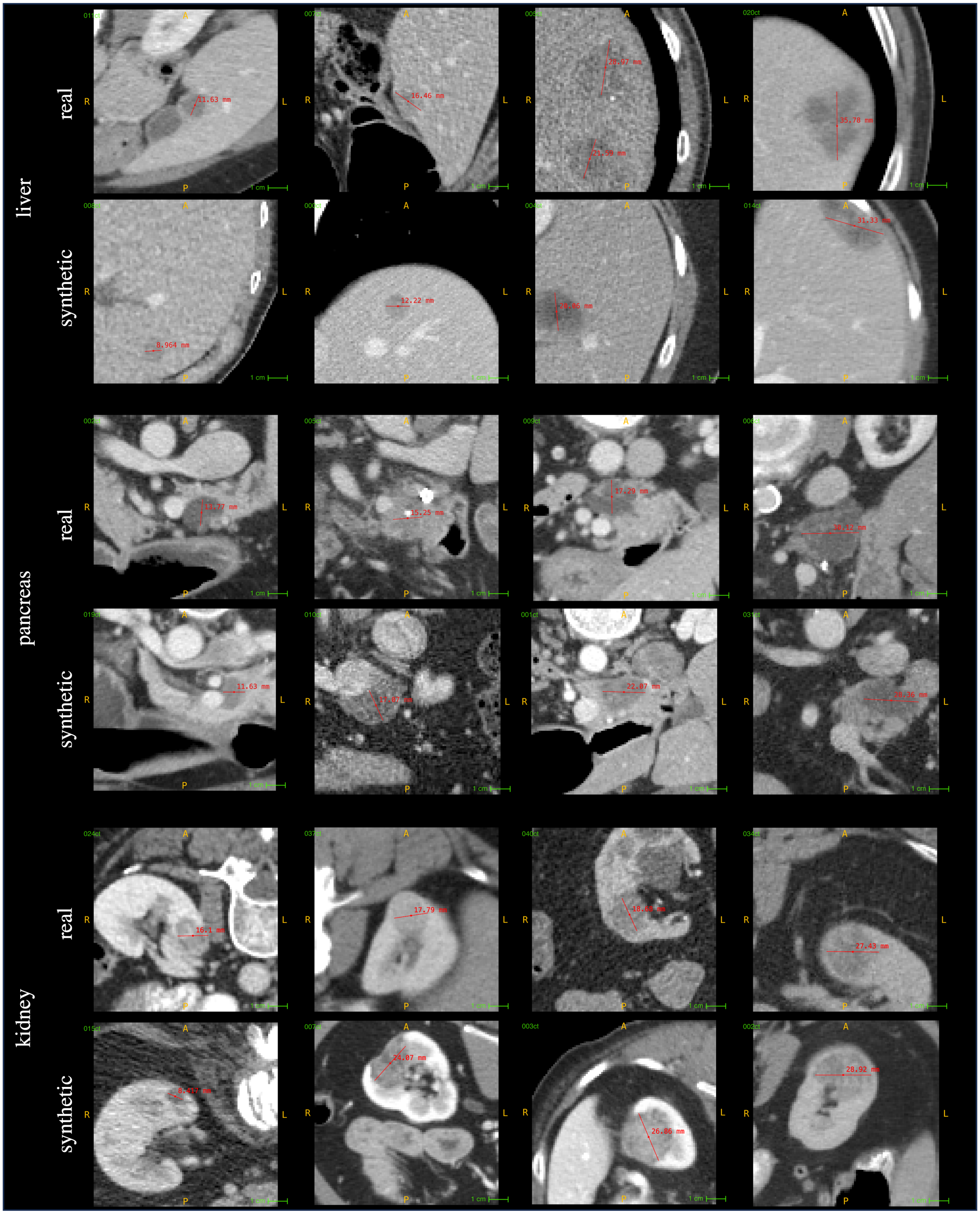


Figure 9. **Visual examples for the Visual Turing Test.** We present real and synthetic tumors, arranged from smaller to larger sizes (columns 1–4), for the Visual Turing Test. The radiologists are instructed to classify each tumor as either real or synthetic. Based on results in §4.1 and Table 1, a minimum of 50% synthetic tumors are identified as real by both radiologists.

B. Description of Radiomics Features

Radiomics Features [76] consist of a comprehensive set of quantitative, high-dimensional imaging attributes derived from radiographic images, such as Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) scans. These attributes capture a broad spectrum of image characteristics, including shape, intensity, texture, and wavelet, among others. The scope of Radiomics Features applications is expansive, ranging from predicting disease prognosis to formulating treatment strategies and evaluating treatment response. The effectiveness of these features has been validated across various medical fields, notably in oncology, neurology, and cardiology.

The key characteristics of Radiomics Features include their high-throughput capacity and reproducibility, which facilitate a detailed characterization of tumor phenotypes. These features are multivariate, incorporating first-order statistics, shape and size-based features, textural features, and filter-based features.

In this paper, we utilize the official Radiomics feature repository[4] to extract the appearance features, which include 3D shape-based features (16 dimensions), gray level co-occurrence matrix (24 dimensions), gray level run length matrix (16 dimensions), gray level size zone matrix (16 dimensions), neighboring gray-tone difference matrix (5 dimensions), and gray level dependence matrix (14 dimensions). The shape descriptors are independent of the gray value and are extracted from the tumor mask. The definitions of these features can be referenced at https://pyradiomics.readthedocs.io/en/latest/features.html. Based on the tumor mask annotations, we are able to extract only the appearance features of tumors. Consequently, for each early-stage tumor, a 91-dimensional vector can be obtained. Ultimately, these features from all early-stage tumors are aggregated for Radiomics feature analysis in Figure 2.

[4] https://github.com/AIM-Harvard/pyradiomics/

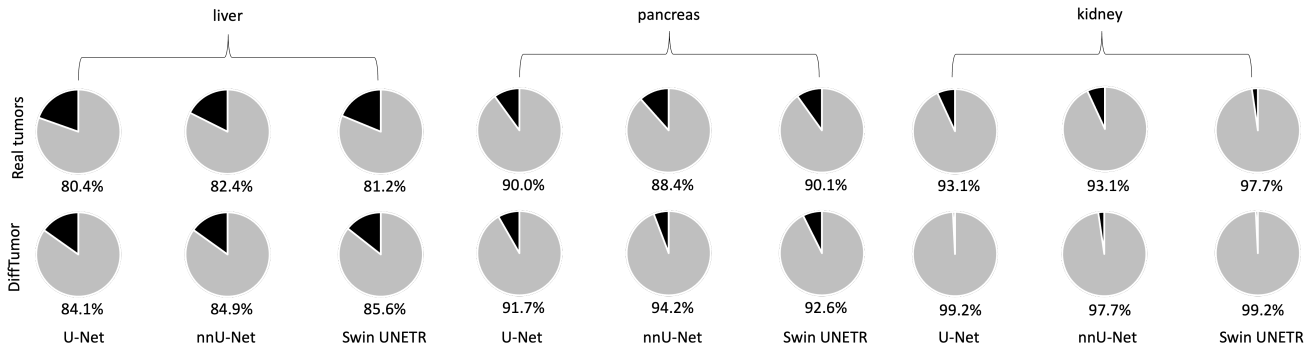


Figure 10. **Enhancement in all-stage tumor detection.** We present the tumor detection results, measured by tumor-wise sensitivity, across different Segmentation Models (U-Net, nnU-Net, Swin UNETR). Consistent with the results in tumor segmentation performance, DiffTumor can significantly enhance tumor detection performance across these three common backbones.

C. Generalizable to Multiple Organs

nnU-Net [39]

source \ target		liver	pancreas	kidneys
liver	real tumors	77.4	1.8	2.4
	Hu *et al*. [37]	78.0	56.3	61.9
	DiffTumor	**80.9**	**60.7**	**71.4**
pancreas	real tumors	1.5	67.0	2.4
	Hu *et al*. [37]	**76.2**	68.8	61.9
	DiffTumor	72.8	**75.0**	**81.0**
kidney	real tumors	0.9	0.9	59.5
	Hu *et al*. [37]	76.2	56.3	69.0
	DiffTumor	**77.4**	**63.4**	**76.2**

Swin UNETR [32]

source \ target		liver	pancreas	kidneys
liver	real tumors	76.2	2.7	0
	Hu *et al*. [37]	79.2	63.4	71.4
	DiffTumor	**83.1**	**69.6**	**73.8**
pancreas	real tumors	1.6	70.5	4.8
	Hu *et al*. [37]	66.4	73.2	71.4
	DiffTumor	**76.2**	**79.5**	**90.5**
kidney	real tumors	0.6	0	69.0
	Hu *et al*. [37]	66.4	**63.4**	76.2
	DiffTumor	**76.2**	61.6	**81.0**

Table 3. **Generalizable across various organs.** We show the comparison of generalization for early-stage tumor detection (measured in tumor-wise Sensitivity %) using additional backbones. The scores highlighted in bold denote the superior performance in each respective domain. Consistent with the results on U-Net presented in Table 2, DiffTumor demonstrates superior performance across nearly all domains on nnU-Net and Swin UNETR.

U-Net	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	62.3	72.8	63.8	54.5	59.0	62.5
	NSD (%)	63.4	74.6	63.3	55.5	61.6	63.7
DiffTumor	DSC (%)	70.9	74.0	67.9	59.1	60.6	66.5
	NSD (%)	71.2	73.9	70.4	61.0	63.3	68.0

nnU-Net	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	64.3	70.2	64.3	56.3	59.6	62.9
	NSD (%)	65.7	72.7	63.1	59.3	62.6	64.7
DiffTumor	DSC (%)	73.6	73.9	67.6	64.9	63.8	68.8
	NSD (%)	75.3	73.9	67.9	69.0	66.5	70.5

Swin UNETR	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	65.1	69.4	57.4	59.0	58.2	61.8
	NSD (%)	65.9	71.7	53.1	62.1	61.9	62.8
DiffTumor	DSC (%)	71.4	71.7	71.6	62.2	62.4	67.9
	NSD (%)	73.5	72.4	74.5	66.5	66.0	70.6

real tumors denotes Segmentation Model trained on 95 CT scans containing real tumors.
DiffTumor denotes Segmentation Model trained on 95 CT scans containing real tumors and 116 healthy CT scans.

Table 4. **Liver tumor segmentation performance on 5-fold cross-validation.** We conduct a comparative analysis of the Segmentation Model (U-Net, nnU-Net, Swin UNETR) trained on both synthetic and real tumors against the model trained exclusively on real tumors, employing 5-fold cross-validation. The evaluation metrics employed include the Dice Similarity Coefficient (DSC) and the Normalized Surface Distance (NSD). DiffTumor consistently enhances liver tumor segmentation performance across these three backbones.

U-Net	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	56.0	51.9	45.5	59.4	43.2	51.2
	NSD (%)	51.0	49.9	43.6	57.7	40.2	48.5
DiffTumor	DSC (%)	64.8	58.0	57.7	67.9	51.8	60.0
	NSD (%)	60.5	55.3	58.3	67.5	47.9	57.9

nnU-Net	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	59.9	50.0	44.8	63.5	50.4	53.7
	NSD (%)	55.7	47.0	47.0	62.3	48.0	52.0
DiffTumor	DSC (%)	63.6	60.5	62.5	67.8	55.3	61.9
	NSD (%)	61.1	59.1	63.4	67.7	54.9	61.2

Swin UNETR	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	52.2	49.5	50.9	60.0	52.1	52.9
	NSD (%)	49.1	49.7	50.9	58.9	47.4	51.2
DiffTumor	DSC (%)	62.2	60.2	59.0	69.7	53.8	61.0
	NSD (%)	58.7	58.4	62.8	67.2	51.2	59.7

real tumors denotes Segmentation Model trained on 96 CT scans containing real tumors.
DiffTumor denotes Segmentation Model trained on 96 CT scans containing real tumors and 120 healthy CT scans.

Table 5. **Pancreatic tumor segmentation performance on 5-fold cross-validation.** We execute a comparative study of the Segmentation Model (U-Net, nnU-Net, Swin UNETR) trained on both synthetic and real tumors against the model trained exclusively on real tumors, utilizing 5-fold cross-validation. The employed evaluation metrics are the Dice Similarity Coefficient (DSC) and the Normalized Surface Distance (NSD). DiffTumorconsistently yields a significant improvement in pancreatic tumor segmentation across these three backbones. It should be noted that the segmentation of pancreatic tumors is deemed the most challenging task among the three abdominal organs in study. The enhancement observed in pancreatic tumor segmentation is the most substantial among the three.

U-Net	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	75.1	68.0	69.0	78.1	70.6	72.0
	NSD (%)	68.4	59.0	57.7	68.3	62.4	63.2
DiffTumor	DSC (%)	84.2	76.7	79.4	80.6	74.1	79.0
	NSD (%)	76.6	64.5	70.7	71.7	65.8	69.9

nnU-Net	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	73.8	76.8	80.0	80.5	73.4	76.9
	NSD (%)	62.7	70.2	71.2	70.8	67.5	68.5
DiffTumor	DSC (%)	84.5	83.4	81.6	83.9	77.3	82.1
	NSD (%)	78.3	74.4	74.1	76.9	72.3	75.2

Swin UNETR	metrics	fold0	fold1	fold2	fold3	fold4	average
real tumors	DSC (%)	80.6	64.9	79.1	76.0	72.2	74.6
	NSD (%)	74.2	55.5	68.2	67.4	64.8	66.0
DiffTumor	DSC (%)	85.1	77.2	81.2	85.7	79.9	81.8
	NSD (%)	79.2	70.8	74.1	78.0	74.1	75.2

real tumors denotes Segmentation Model trained on 96 CT scans containing real tumors.
DiffTumor denotes Segmentation Model trained on 96 CT scans containing real tumors and 120 healthy CT scans.

Table 6. **Kidney tumor segmentation performance on 5-fold cross-validation.** We perform a comparative analysis of the Segmentation Model (U-Net, nnU-Net, Swin UNETR) trained on both synthetic and real tumors versus the model trained exclusively on real tumors, employing 5-fold cross-validation. Our evaluation metrics include the Dice Similarity Coefficient (DSC) and the Normalized Surface Distance (NSD). Similar to the other two tumor segmentation tasks, DiffTumor can deliver substantial improvements in kidney tumor segmentation across these three prevalent backbones.

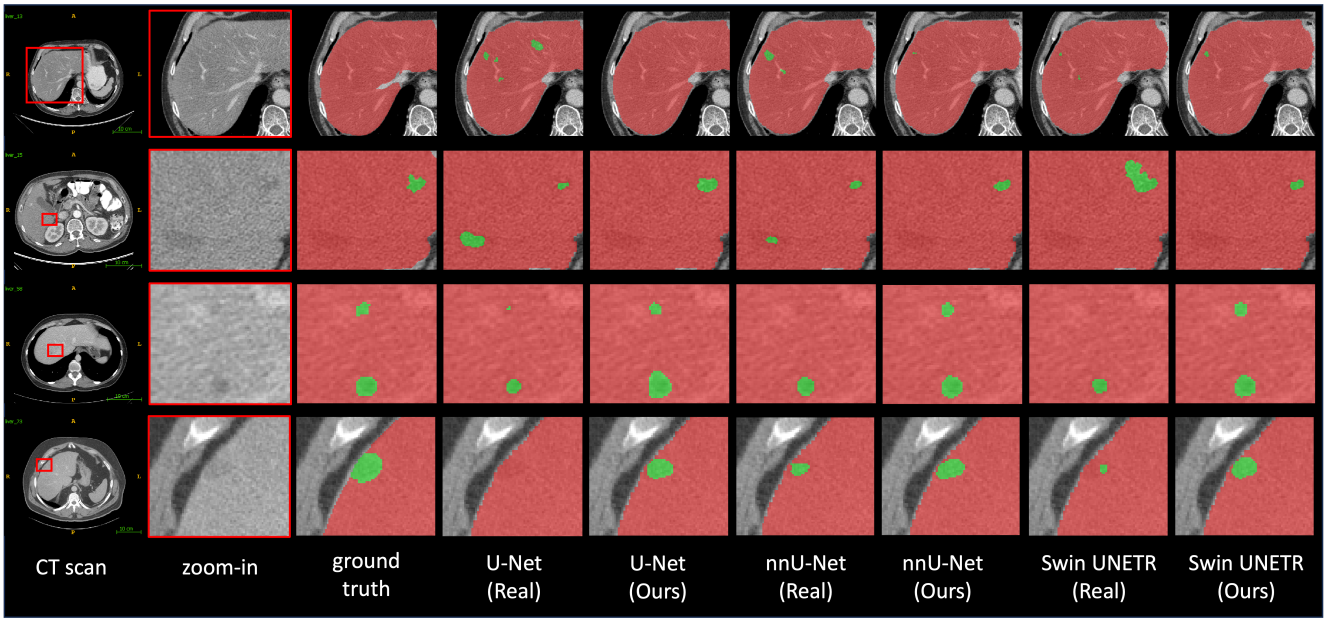


Figure 11. **Liver tumor segmentation.** We provide qualitative visualizations of Segmentation Models (U-Net, nnU-Net, Swin UNETR) for liver tumor segmentation. The results from the first and second rows illustrate that DiffTumor can effectively reduce the number of false positive cases to a certain extent. The results in the third and fourth rows indicate that DiffTumor can improve the detection rate within the tumor region. Consequently, our method yields a substantial improvement in segmentation performance, as evidenced in Table 4.

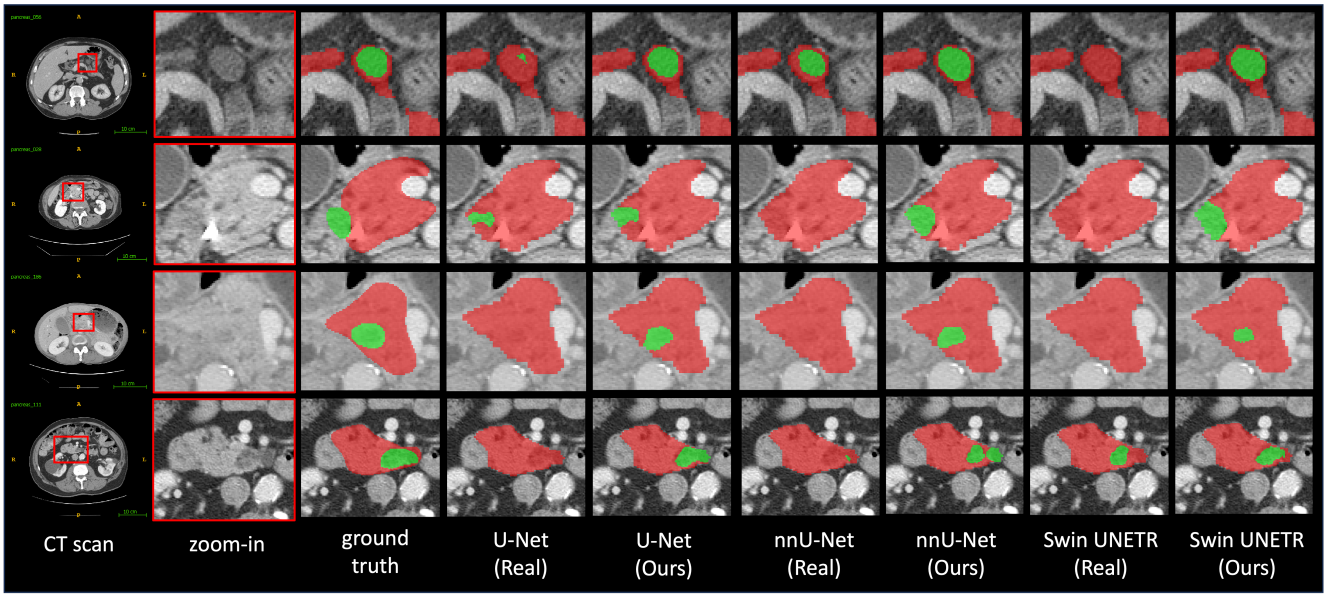


Figure 12. **Pancreatic tumor segmentation.** We provide qualitative visualizations of Segmentation Models (U-Net, nnU-Net, Swin UNETR) for pancreatic tumor segmentation. Pancreatic tumor segmentation is more challenging as many tumors are easily missed by Segmentation Models. The results displayed in rows 1-4 indicate that DiffTumor can aid detecting tumors that were missed during training on real tumors, thereby resulting in a significant improvement in segmentation performance, as illustrated in Table 5.

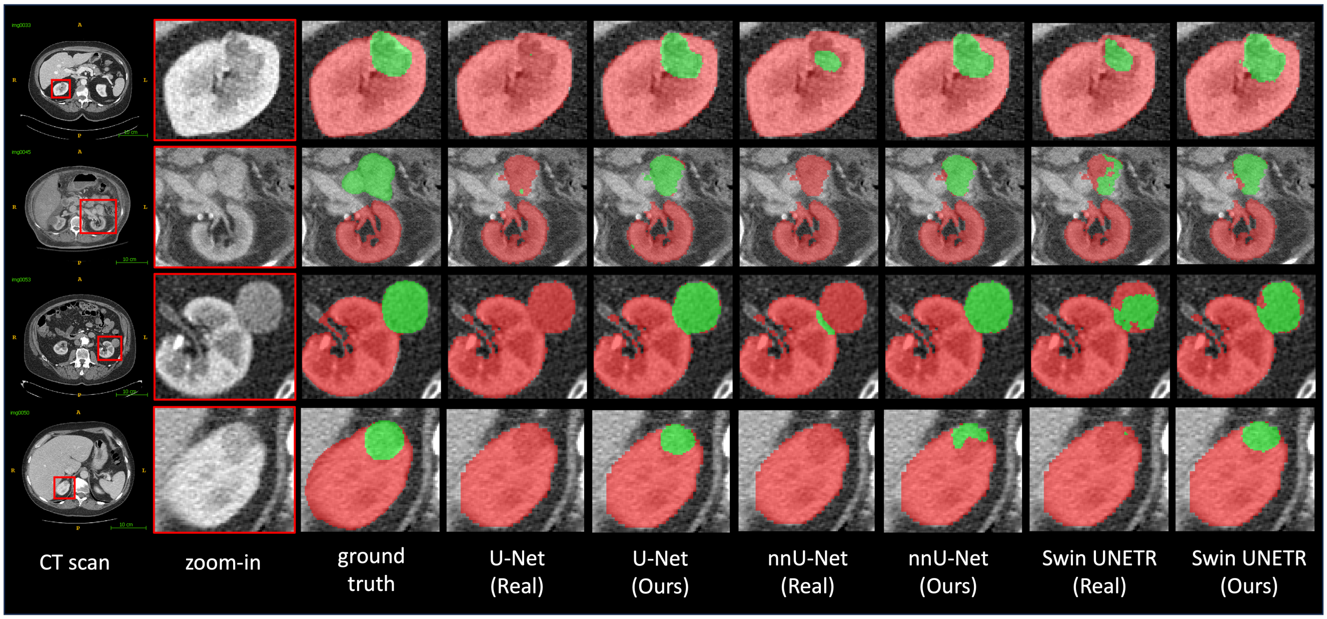


Figure 13. **Kidney tumor segmentation.** We provide qualitative visualizations of Segmentation Models (U-Net, nnU-Net, Swin UNETR) for kidney tumor segmentation. Segmentation Models, when trained on real tumors, will miss the tumors located on the boundary of the kidney or outside the kidney. DiffTumor can help enhance the segmentation of these difficult cases, thereby improving the overall segmentation performance, as evidenced in Table 6.

D. Generalizable to Different Patient Demographics

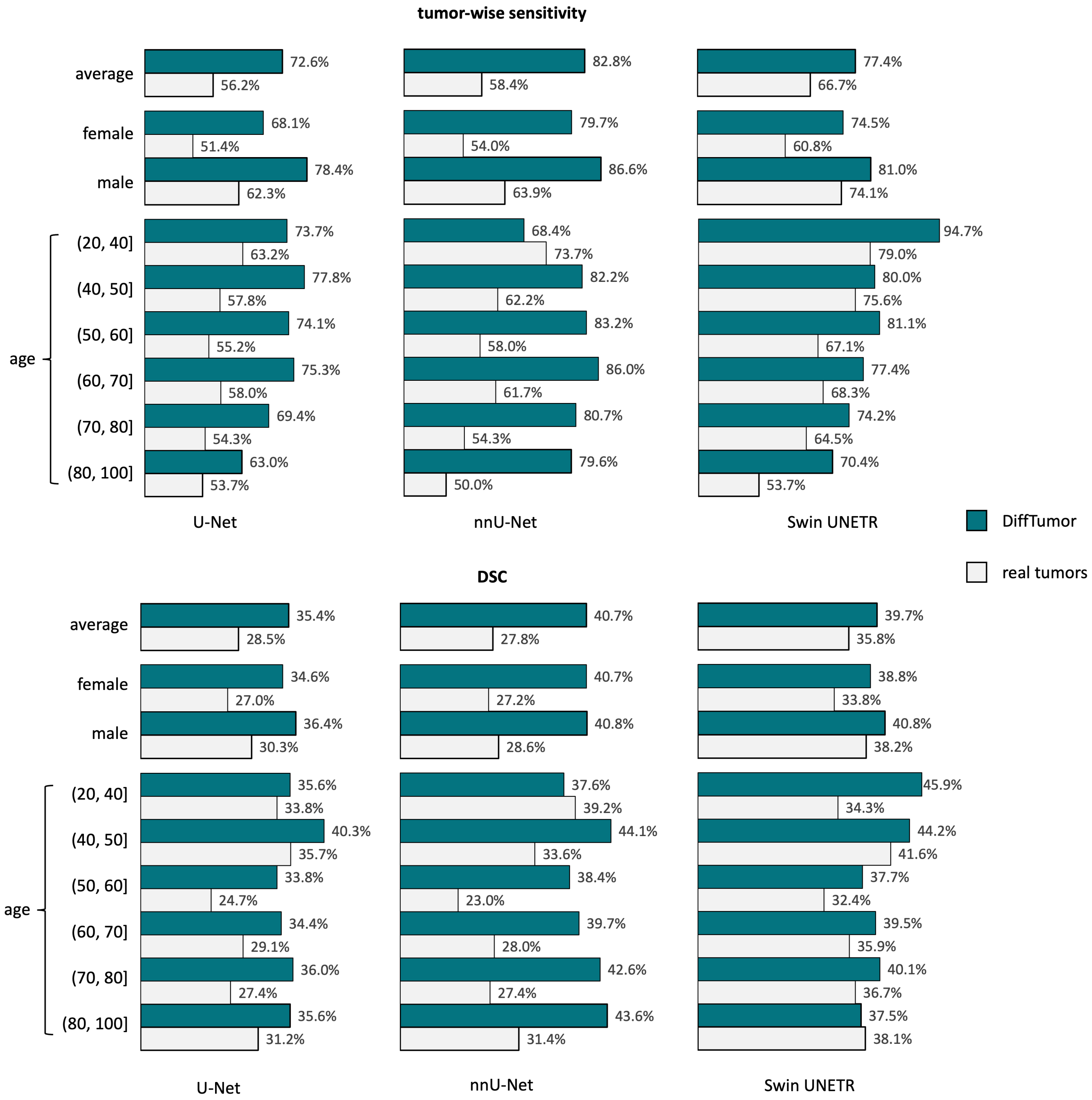


Figure 14. **Generalizable across different patient demographics.** A comparison of generalization across different patient demographics for tumor detection (tumor-wise Sensitivity %) and segmentation (DSC %) using different backbones is provided. DiffTumor can consistently enhance tumor detection and segmentation performance by a significant margin across various patient groups and different backbones. The most significant improvement is observed on nnU-Net.

E. Dataset & Implementation Details

E.1. Dataset Details

Real-tumor datasets. LiTS [6] comprises 131 and 70 contrast-enhanced 3-D abdominal CT scans for training and testing, respectively. This dataset was compiled utilizing various scanners and protocols from six unique clinical sites, which resulted in a significant variation in in-plane resolution (ranging from 0.55 to 1.0 mm) and slice spacing (ranging from 0.45 to 6.0 mm). The MSD-Pancreas [2] dataset comprises 420 portal-venous phase CT scans from patients who underwent pancreatic mass resection. It includes 281 CT scans designated for training and 139 CT scans for testing. The annotations provided correspond to the pancreatic parenchyma and pancreatic mass. KiTS [33] includes 210 CT scans for training and 90 CT scans for testing. Each CT scan features one or more kidney tumors. The University of Minnesota Medical Center provides the annotations.

Healthy-organ datasets. AbdomenAtlas-8K [64] is currently the most extensive multi-organ dataset, with annotations for the spleen, liver, kidneys, stomach, gallbladder, pancreas, aorta, and IVC in 8,448 CT volumes, which equates to 3.2 million slices. AbdomenAtlas-8K consolidates datasets from 26 distinct hospitals worldwide. In this study, we utilize the CLIP-Driven Universal Model [55, 92] to select CT scans that feature the corresponding healthy abdominal organs. This model, ranking first in the Medical Segmentation Decathlon (MSD) competition, has demonstrated high sensitivity and specificity in tumor detection. Consequently, we employ the pre-trained weights[5] using the Swin UNETR backbone to identify CT scans where the prediction includes the organs but does not include the corresponding tumors. Through this process, we have obtained 1246 CT volumes with healthy livers, 1901 CT volumes with healthy pancreas, and 1005 CT volumes with healthy kidneys.

Proprietary dataset. It comprises 5,038 CT scans with 21 annotated organs, with each case having been scanned by contrast-enhanced CT in both venous and arterial phases, utilizing Siemens MDCT scanners. In this study, we utilize 532 CT scans containing 690 PDAC to assess the generalizability of the Segmentation Model across varied patient demographics. The test set we used includes 243 CT scans of males and 289 CT scans of females. Additionally, the age range of these patients spans from 20 to 146, essentially covering all stages of a person's life.

E.2. Implementation Details

Autoencoder Model. In this study, we train Autoencoder Model on a total of 9262 CT scans from the AbdomenAtlas-8K dataset and a private dataset. The purpose is to learn a general low-dimensional latent representation of CT scans. The model processes the CT volume into a latent feature, reducing the original input volume's size by 1/4 in height, width, and depth, respectively. We set the codebook size and dimensionality at 16384 and 8, respectively. CT scan orientation is adjusted to specific axcodes, and isotropic spacing is applied to resample each scan, resulting in a uniform voxel size of $1.0 \times 1.0 \times 1.0mm^3$. Additionally, the intensity in each scan is truncated to the range $[-175, 250]$ and then linearly normalized to $[-1, 1]$. During training, we crop random fixed-sized $96 \times 96 \times 96$ regions. We employ the Adam optimizer for training with β_1 and β_2 hyperparameters set to 0.9 and 0.999, respectively, a learning rate of 0.0003, and a batch size of 4 per GPU. The training is conducted over a week on a node with four A100 GPUs, completing 200k iterations.

Diffusion Model. In this study, we train the corresponding Diffusion Model specifically for tumors of three different abdominal organs. The data preprocessing carried out during the training phase is identical to the approach used for training Autoencoder Model. Besides, we utilize the Adam optimizer for training with β_1 and β_2 hyperparameters set to 0.9 and 0.999, respectively, a learning rate of 0.0001, and a batch size of 10 per GPU. The training is conducted over the course of a day on a node with an A100 GPU for 60k iterations.

Segmentation Model. The code for the Segmentation Model is implemented in Python using MONAI[6]. In this study, we implement Swin UNETR based on the Swin UNETR Base variant. The orientation of CT scans is adjusted to specific axcodes. Isotropic spacing is utilized to resample each scan to achieve a uniform voxel size of $1.0 \times 1.0 \times 1.0mm^3$. Besides, the intensity in each scan is truncated to the range $[-175, 250]$ and then linearly normalized to $[0, 1]$. During training, we crop random fixed-sized $96 \times 96 \times 96$ regions with the center being a foreground or background voxel based on the predefined ratio. Additionally, the input patch is randomly rotated by 90 degrees, and the intensity is shifted with a 0.1 offset, each with probabilities of 0.1 and 0.2, respectively. To avoid confusion between the organs on the right and left sides, mirroring augmentation is not employed. All models on real tumors are trained for 3,000 epochs and models on synthetic and real

[5] https://github.com/ljwztc/CLIP-Driven-Universal-Model

[6] https://monai.io/

tumors are trained for 2,000 epochs. Moreover, the base learning rate is set at 0.0002, and the batch size is set at two. We adopt the linear warmup strategy and the cosine annealing learning rate schedule. For details on the tumor synthesis process during the training of the Segmentation Model, please refer to the provided code. For inference, we use the sliding window strategy by setting the overlapping area ratio to 0.75. Besides, to rule out tumor mask predictions that do not belong to the respective organs, we use the pseudo labels of organs obtained through [55] to process the predictions of Segmentation Models.

F. Discussion

F.1. Comparison with Related Works

In recent works, Hu *et al*. [37] have synthesized tumors in the liver using a model-based approach. This approach, guided by radiologists, involves several image-processing operations such as ellipse generation, elastic deformation, salt-noise generation, Gaussian filtering, scaling, and clipping. The synthetic tumors are realistic in comparison to real liver tumors. Notably, the AI trained with synthetic tumors achieves segmentation/detection performance that is comparable to the performance of the AI trained with real tumors.

However, the approach of Hu *et al*. [37] requires significant effort and expertise to identify the proper imaging characteristics of tumors. In other words, the resulting synthetic tumors need to be explicitly specified by radiologists, tailored to the specific types of tumors and must be redesigned for tumors in other organs. To demonstrate the superiority of DiffTumor for enhancing tumor segmentation in the three abdominal organs, we compare it with the representative tumor synthetic strategy by Hu *et al*. [37]. The results can be found in Table 7.

Methods	liver		pancreas		kidneys	
	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
real tumors	62.3	63.4	56.0	51.0	75.1	68.4
Hu *et al*. [37]	69.7	70.9	55.9	49.9	80.8	71.0
DiffTumor	**70.9**	**71.2**	**64.8**	**60.5**	**84.2**	**76.6**

Table 7. **Comparison for tumor segmentation enhancement.** The comparison for all-stage tumor segmentation is conducted based on the U-Net backbone. While Hu *et al*. [37] is designed for liver tumor synthesis, DiffTumor can bring more significant improvement in liver tumor segmentation. The synthesized tumors by Hu *et al*. [37] can also boost the DSC and NSD scores for kidney tumor segmentation. However, DiffTumor can yield better results. Additionally, when adding the synthesized tumors by Hu*et al*. [37] for pancreatic tumor segmentation, the DSC and NSD scores even drop compared with training solely on real tumors. This suggests that the synthetic strategy may not be suitable for the synthesis of pancreatic tumors. On the contrary, DiffTumor can significantly improve pancreatic tumor segmentation. These results underline the superiority of DiffTumor in enhancing tumor segmentation across various abdominal organs.

F.2. Unrealistic Generation

Although DiffTumor is capable of generating highly realistic tumors, it also occasionally produces some that are less convincing. Consequently, about 50% of the tumors are identified as inauthentic by the more experienced radiologist in the Visual Turing test. The tumors deemed inauthentic fall short in several aspects, such as shape, attenuation and noise distribution. Some synthetic tumors have inaccurate shapes, resembling flat, strip-like, or sickle-shaped lesions. In contrast, early-stage tumors originating from parenchymal organs typically exhibit a round or oval shape. Larger tumors fail to display a mass effect, characterized by the displacement of normal structures due to the tumor's inherent volume. Furthermore, some synthetic tumors inaccurately display a lower density, which is similar to that of fat or fluid. Finally, the noise distribution in some synthetic tumors does not match that in the CT background. We show several unrealistic generation cases in Figure 15.

F.3. Challenging Cases Analysis

Instances of low performance are observed with all Segmentation Models (U-Net, nnU-Net, Swin UNETR) trained on both synthetic and real tumors. These instances often involve tumors with uncommon imaging features, as identified by experienced radiologists. Examples of these tumors from the liver, kidney, and pancreas are provided in Figure 16.

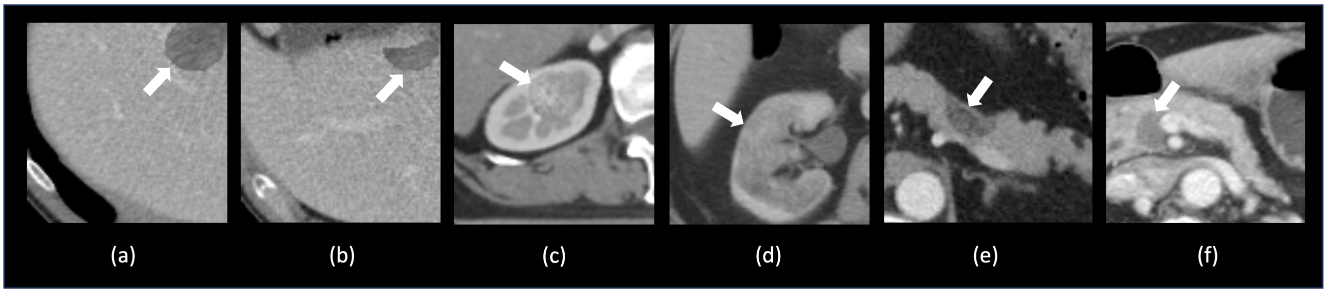


Figure 15. **Unrealistic generation cases.** **(a)** A synthetic liver tumor with an edge that is too sharp for a malignant tumor, and the inner noise has a discrepancy from the surrounding tissue. **(b)** A synthetic tumor with a sickle shape, which is unrealistic for a liver tumor. **(c)** A synthetic kidney tumor in the corticomedullary junction zone exhibiting a round nodular shape. However, this lesion fails to display a mass effect, as the healthy surrounding renal structure shows no deformation due to the tumor's inherent volume. Additionally, the texture and noise inside the tumor do not match the CT background. **(d)** A synthetic kidney tumor has a shape that matches the kidney "perfectly." However, solid tumors tend to grow expansively into a round nodular shape rather than precisely overlapping with the organ. **(e)** A synthetic tumor with a sickle shape, which is unrealistic for a pancreatic tumor. Additionally, the attenuation of this lesion is too low for a solid pancreatic tumor. **(f)** A synthetic pancreatic tumor adjacent to extra-pancreatic vessels. This tumor shows no mass effect, leaving the vessels without displacement or infiltration.

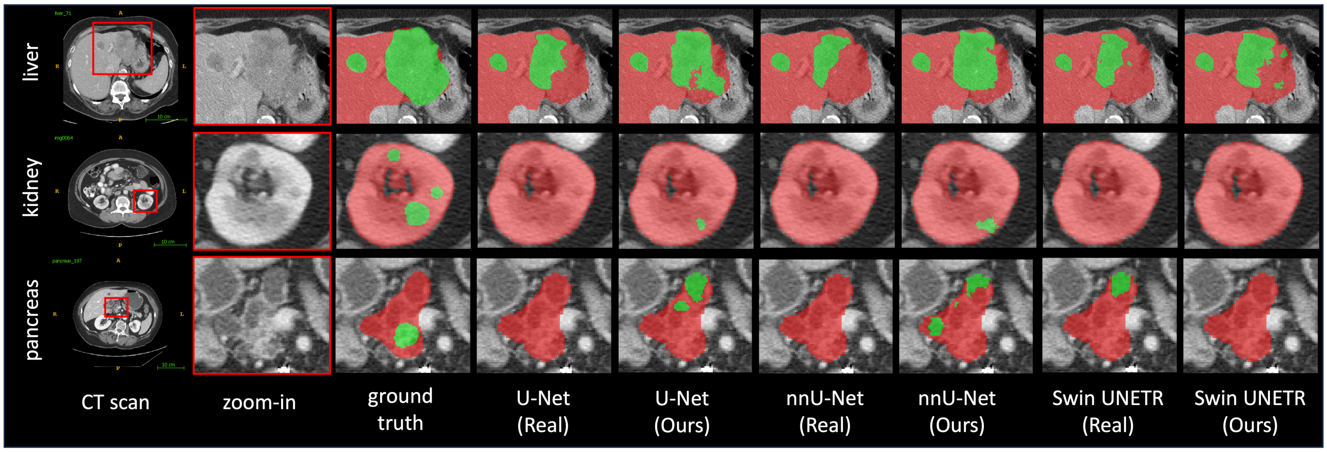


Figure 16. **Challenging cases for tumor segmentation. Liver tumor:** An irregular heterogeneous hypoattenuating mass is observed in the left lateral lobe of the liver, showing extracapsular penetration and infiltration into the stomach. The area with lower hypoattenuation inside the tumor indicates necrosis. **Kidney tumor:** This is a case with multiple renal tumors in the cortex and corticomedullary junction zone. Some of the tumors exhibit isoattenuation with the renal medulla, posing difficulties in detecting the lesions. **Pancreatic tumor:** In the illustrated case within this figure, the segmentation model erroneously classifies dilated pancreatic ducts as a tumor, resulting in inaccurate segmentation. Conversely, the genuine tumor, characterized by its indistinct boundaries, is not successfully segmented.