Towards Generalizable Tumor Synthesis

Qi Chen¹  Xiaoxi Chen²  Haorui Song³  Zhiwei Xiong¹
Alan Yuille³  Chen Wei³  Zongwei Zhou³,*

¹University of Science and Technology of China
²Shanghai Jiao Tong University
³Johns Hopkins University

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

Abstract

Tumor synthesis enables the creation of artificial tumors in medical images, facilitating the training of AI models for tumor detection and segmentation. However, success in tumor synthesis hinges on creating visually realistic tumors that are generalizable across multiple organs and, furthermore, the resulting AI models being capable of detecting real tumors in images sourced from different domains (e.g., hospitals). This paper made a progressive stride toward generalizable tumor synthesis by leveraging a critical observation: early-stage tumors (< 2cm) tend to have similar imaging characteristics in computed tomography (CT), whether they originate in the liver, pancreas, or kidneys. We have ascertained that generative AI models, e.g., Diffusion Models, can create realistic tumors generalized to a range of organs even when trained on a limited number of tumor examples from only one organ. Moreover, we have shown that AI models trained on these synthetic tumors can be generalized to detect and segment real tumors from CT volumes, encompassing a broad spectrum of patient demographics, imaging protocols, and healthcare facilities.

1. Introduction

Tumor synthesis enables the creation of artificial tumor examples in medical images [11, 42, 88], it is particularly valuable when there is a dearth or complete absence of per-voxel annotated real tumors (e.g., early-stage tumors) for effective AI training. Typically, to train AI models for tumor detection in multiple (N) organs, annotated real tumor examples from each of these organs are necessary, and ideally, in substantial numbers [14, 43, 54, 55, 97]. Furthermore, AI models often fail to generalize across images from different hospitals, which may vary due to variations in imaging protocols, patient demographics, and scanner manufacturers [61, 95, 96]. The challenge amplifies with the need for extensive manual annotations, a task that could demand up to 25 human years for annotating just one tumor type [1, 9, 83]. The task of collecting and annotating a comprehensive dataset encompassing tumors from multiple organs (N) and images from numerous hospitals (M) is daunting, considering both annotation cost and complexity (N × M). We hypothesize that tumor synthesis could solve this challenge by creating various tumor types across non-tumor images from multiple hospitals, even when only one tumor type is available, thereby simplifying the complexity from N × M to 1 × M.

Success in tumor synthesis hinges on creating visually realistic tumors that are generalizable across multiple organs and, furthermore, the resulting AI models being generalizable in detecting real tumors in images sourced from different hospitals. Previous studies have introduced generative models to create synthetic medical data (not limited to tumors) such as polyp detection from colonoscopy videos [69], COVID-19 detection from Chest X-ray [26, 56, 87], and diabetic lesion detection from retinal images [78]—refer to §5 for a comprehensive review. However, these studies have primarily focused on enhancing the detection and segmentation of specific tumors without fully exploring the wider generalizability of these models across different organs and patient demographics.

This paper made a progressive stride toward generalizable tumor synthesis by leveraging a critical observation: early-stage tumors (< 2cm) tend to have similar imaging characteristics in computed tomography (CT)¹. Early-stage tumors typically present small, round, or oval shapes with minimal deformation and exhibit relatively simple and uni-

¹Note that, owing to the public dataset constraints, we have only verified the similarity across early hepatocellular carcinoma and intrahepatic cholangiocarcinoma from the liver, pancreatic ductal adenocarcinoma from the pancreas, and renal cell carcinoma from kidneys.
Figure 1. Generalizable tumor synthesis across organs. Early-stage tumors present similar imaging characteristics in computed tomography (CT), whether they are located in the liver, pancreas, or kidneys. Leveraging this observation, we develop a generative AI model on a few examples of annotated tumors in a specific organ, e.g., the liver (in purple). This AI model (in purple), trained exclusively on liver tumors, can directly create synthetic tumors in those organs where CT volumes of annotated tumors are relatively scarce, e.g., the pancreas (in cyan) and kidneys (in blue and green). By integrating synthetic tumors into extensive CT volumes of healthy organs—routinely collected in clinical settings—we can substantially augment the training set for tumor segmentation. This enhancement can also significantly improve the AI generalizability across CT volumes sourced from diverse hospitals and patient demographics.

form textures in CT volumes [13]. Hence, early tumors in parenchymal organs (e.g., liver, spleen, pancreas, adrenal glands, and kidneys) should appear similarly, as shown in Figure 1. The major difference is the contrast between the tumors and background organs or other anatomical structures rather than the tumors themselves. Using four public datasets and our proprietary datasets, §2 verifies the similarity of early-stage tumors across various organs.

Leveraging this observation, we introduce a novel framework, termed DiffTumor, that can learn the common imaging characteristics of tumors across various organs, and the generated synthetic tumors are useful for training AI models to detect and segment real tumors from CT volumes of varying patient demographics. The development of DiffTumor is composed of three stages. ① Training an Autoencoder Model—consisting of an encoder and decoder—on 9,262 unlabeled three-dimensional CT volumes. The use of large, diverse datasets can enhance the model’s ability to generalize across CT volumes of different patient demographics and reduce the need for annotated tumor volumes for training Diffusion Models in the subsequent stages. The proxy task is image reconstruction, which facilitates the model in learning comprehensive latent features. ② Training a Diffusion Model—a specific type of generative models—using latent features and tumor masks as conditions. Once trained, this model can generate latent features necessary for reconstructing CT volumes with tumors based on arbitrary masks. ③ Training a Segmentation Model using synthetic tumors, which are reconstructed by the decoder, and their corresponding masks. With a large repository of healthy CT volumes, our DiffTumor framework can produce a vast array of synthetic tumors, varying in location, size, shape, texture, and intensity, therefore fostering high-performing AI models for tumor detection/segmentation.

The key contributions of this paper are two-fold. Firstly, we have verified with feature analysis, reader studies, and clinical knowledge that early-stage tumors (<2cm) manifest with similar imaging characteristics across various organs in CT volumes, establishing the foundation for the development of generalizable tumor synthesis. Secondly, we have developed a three-stage tumor synthesis framework, DiffTumor, that trains generative models with minimal annotations (Figure 5; one annotated CT volume), creates synthetic tumors in real-time (Figure 6; 100 ms/tumor), and improves early-stage tumor detection (Figure 7; improved sensitivity up to +28.6%). In summary, compared with training AI on extensively annotated CT volumes of real tumors, our DiffTumor is generalizable from two critical perspectives.

1. DiffTumor can create visually realistic tumors generalizable to a range of organs even when the diffusion model was trained on a limited number of tumor examples from a specific organ (§4.2; +10.7% DSC).
2. DiffTumor can develop an AI model to detect and segment real tumors generalizable to a variety of CT volumes of varied patient demographics, imaging protocols, and healthcare facilities (§4.3; +9.1% DSC).

2. Preliminary

We observe that early-stage tumors (<2cm)\(^2\) often share similar imaging characteristics in CT volumes, whether

\(^2\)Based on the TNM system, the most widely used staging system for classifying a malignancy tumor, we recognize a primary malignant tumor with a diameter less than 2 cm and no evidence of nearby lymph node involvement or metastasis as an early-stage tumor [8, 24, 66].
we assess the performance of a support vector machine (SVM) classifier, using Radiomics features [16], and three expert radiologists in identifying the originating organs of cropped tumors. The SVM classifier is tasked with a three-way classification to ascertain whether a tumor originates from the liver, pancreas, or kidneys. In a similar test, radiologists examine the original CT images of these tumors. Reader study results on the left panel indicate significant challenges for both the SVM classifier and the radiologists in accurately identifying the origin of early-stage tumors. The precision and recall scores for both methods closely resemble those of random guessing. Additionally, on the right panel, we present a t-SNE visualization of Radiomics features for tumors from the liver, pancreas, and kidneys. These results highlight the considerable similarity in features and images of early-stage tumors.

they originate in the liver, pancreas, or kidneys. This observation, when confirmed, could have profound implications for generative AI in medical imaging. It suggests that generative AI might be trained on one tumor type, for which data and annotations are more easily obtained, and then applied to create various tumor types in different organs, where acquiring sufficient data can be challenging. Using synthetic tumors can substantially improve AI performance in tumor detection and segmentation in practice. In light of this, we have rigorously pursued the validation through four approaches as follows.

(1) Radiologist reader study. The objective of the reader study is to assess the ability of radiologists to recognize the organ class of early cancer. We uniformly crop 360 CT images of the tumor region from three abdominal organs, as per the annotations. In order to exclude the influence of surrounding organ textures on recognition, we only retain a small amount of organ textures in the tumor boundary region. Examples of CT crops used for the reader study are provided in Figure 2, and more examples are in Appendix A. Three expert radiologists, qualified under the Quality Standards Act, participate in the reader study. The recognition results are shown in Figure 2(a). The nearly random probability of the precision and recall scores indicates that the appearance of early-stage tumors is so similar that even experienced radiologists have difficulty distinguishing the organ types of these tumors.

(2) Radiomics feature analysis. We now analyze the similarity in Radiomics features of early-stage tumors. Quantitatively, we train three types of learning-based classifiers, including support vector machine (SVM), Random Forests, and AdaBoost, to identify the organ types of early-stage tumors. To draw a general conclusion, we conducted ten repeated experiments and calculated the precision and recall scores of these classifiers in both the training and test sets. The final results for SVM show that the precision and recall for the training set are close to 1, indicating that SVM is well-trained and capable of learning a decent decision boundary for the training set. However, the precision scores for the test set are nearly equivalent to random probability, as shown in Figure 2(a). Similarly, Random Forest achieves a precision of 50.3% and a recall of 54.9%, while AdaBoost achieves a precision of 35.4% and a recall of 49.5%. This suggests that even a well-trained classifier struggles to recognize the organ types of unseen early-stage tumors. Qualitatively, Figure 2(b) visualizes the feature mapping in a two-dimensional space using t-SNE. The appearance features of early-stage tumors are distributed in a joint feature space, and there is no separation for different organ types.

(3) Deep feature analysis. We investigate the similarity of early tumors across different organs using deep features extracted by ResNet and DenseNet. These two networks are trained to classify the types of organs affected by early-stage tumors. ResNet achieves a precision of 59.7% and a recall of 55.6%; DenseNet achieves a precision of 44.3% and a recall of 61.1%. As seen, no matter whether using hand-craft features or deep features, the results reached a consistent
observation—none of the classifiers can distinguish early tumors correctly among the three organs.

(4) Clinical evidence and justification. Tumorigenesis is a gradual, multi-step process involving cellular and histological changes that culminate in successively malignant lesions [13]. Histologically, early-stage tumors often exhibit well-to-moderately differentiated neoplastic cells with mild atypia, limited hemorrhage, and necrosis [4, 15, 20]. This cellular similarity leads to shared imaging features across various parenchymal organs (e.g., liver, pancreas, spleen, adrenal gland, kidney). Early-stage tumors typically appear as relatively homogeneous nodules with indistinct margins and small diameters in CT images [71]. These consistent characteristics, observed across populations, ages, and genders, suggest that tumor synthesis models could learn and generalize shared imaging features across organs.

• Liver tumors: Hepatocellular carcinoma (HCC) in the early stage presents as a small, well-differentiated nodule with minimal metastatic potential [25]. Active neangiogenesis leads to reduced portal triads and an isosattenuating or hypoattenuating appearance (wash-out) in the venous phase compared to surrounding parenchyma [4].

• Pancreatic tumors: Multiphase CT with intravenous contrast is preferred for diagnosing suspected pancreatic lesions [50]. Most pancreatic ductal adenocarcinomas (PDACs) exhibit hypoenhancement relative to surrounding tissue due to their dense fibroblastic stroma [21].

• Kidney tumors: CT is the gold standard for evaluating renal cell carcinoma (RCC) [52]. Clear cell RCC, the most common subtype, typically presents as a small, hypoenhancing renal lesion with surrounding homogenous enhancement in the nephrographic phase [79].

3. DiffTumor

3.1. Autoencoder Model

Diffusion models directly applied to three-dimensional CT volumes incur significant computational costs. To address this, Latent Diffusion Models (LDMs) [67] operate within a compressed, lower-dimensional latent space. Following this approach, we construct our diffusion model within the latent space of 3D CT volumes. Our first step involves training a 3D autoencoder to learn meaningful, compressed latent representations. We adapt the Vector Quantized Generative Adversarial Networks (VQGAN) [22] architecture, replacing 2D convolutions with their 3D counterparts.

Formally, we denote a CT sub-volume as \( x \in \mathbb{R}^{H \times W \times D} \), where \( H \) denotes the height, \( W \) the width, and \( D \) the depth. The CT sub-volume \( x \) is first converted to latent features \( z \) by an encoder \( f \) and a quantization operation \( q \), i.e., \( z = q(f(x)) \in \mathbb{Z}^{h \times w \times d} \), where \( h \) denotes the feature height, \( w \) the feature width, and \( d \) the feature depth. In the vector quantization step, the latent features \( z \) are quantized into \( c_z \in \mathbb{R}^{h \times w \times d \times c} \) by replacing each one with its closest corresponding codebook vector in the learned codebook \( C = \{ c_i \}_{i=1}^{K} \). \( K \) is the codebook size. Finally, a decoder \( g \) reconstructs the latent from \( c_z \) to \( \hat{x} = g(c_z) \). The loss is a summation of three terms:

\[
L_{\text{recon}} + L_{\text{codebook}} + \alpha L_{\text{commit}},
\]

where \( L_{\text{recon}} = \| x - \hat{x} \|_1 \), \( L_{\text{codebook}} = \| \text{sg} [ f(x) ] - c_z \|_2^2 \), and \( L_{\text{commit}} = \alpha \| \text{sg} [ c_i ] - f(x) \|_2^2 \). \( \text{sg} [ \cdot ] \) denotes the stop-gradient operation and \( \alpha \) the coefficient.

In addition to these three loss terms, we also adopt a perceptual loss and a discriminator to improve the reconstruction quality. For 3D CT reconstruction, we adopt a 3D volume discriminator \( D_v \) to penalize implausible artifacts for the 3D reconstruction of \( x \), and a 2D slice discriminator \( D_s \) to encourage per-slice quality. To stabilize the GAN training, we add the feature matching losses \( L_{\text{match}} \). Moreover, due to the CT volumes being preprocessed to isotropic volume, we constrain the high-frequency texture for all three planes of \( x \) by using perceptual loss for projected reconstruction slices \( \hat{x}_{HW} \in \mathbb{R}^{H \times W}, \hat{x}_{HD} \in \mathbb{R}^{H \times D}, \hat{x}_{WD} \in \mathbb{R}^{W \times D} \). The overall objective of the Autoencoder is:

\[
\min_{f, g, k} \left( L_{\text{recon}} + L_{\text{codebook}} + \alpha L_{\text{commit}} + L_{\text{match}} + L_{\text{perceptual}} \right) + \min_{f, g, k} \left( \max_{D_v, D_s} (L_{\text{disc}}) \right),
\]

where

\[
L_{\text{disc}} = \log D_{s/v}(x) + \log (1 - D_{s/v}(\hat{x})).
\]

\[
L_{\text{match}} = \sum_i \| D_{s/v}^{(i)}(\hat{x}) - D_{s/v}^{(i)}(x) \|_2.
\]

\( D_{s/v}^{(i)} \) denotes the \( i \)th layer of discriminators.

3.2. Diffusion Model

We aim to synthesize realistic and diverse CT volumes with tumors to facilitate the training of the tumor segmentation model. Given the fact that healthy CT volumes are much more accessible than CT volumes with tumors, we focus only on tumor synthesis, and we do not intend to model organ textures outside of the tumors, which can be easily obtained from healthy CT volumes. To be specific, our diffusion model is conditioned on both a tumor mask that indicates the shape and location of tumors in the latent feature and the healthy region of CT volumes.

Formally, given a pair of tumor-present CT volume \( x_0 \) and the mask of its tumor region \( m \), the diffusion model is conditioned on both the tumor mask \( m \) and the healthy region \( z_{0, \text{healthy}} := (1 - m) \odot x_0 \). The diffusion model approximates the distribution of the latent features of tumor-present CT volumes. In the forward process, the latent feature \( z_0 \) is gradually converted to white Gaussian noise \( z_T \sim \mathcal{N}(0, 1) \) by recursively adding a small amount of Gaussian noise \( T \) times following the Markov process below:

\[
p(z_{t} | z_{t-1}) = \mathcal{N}(z_{t} \mid \sqrt{1 - \beta_t} z_{t-1}, \beta_t I),
\]
Figure 3. **Overview of the DiffTumor framework.** Towards generalizable tumor synthesis, developing our DiffTumor involves three stages.① Training an *Autoencoder Model*—consisting of an encoder and decoder—to learn comprehensive latent features. The learning task here is image reconstruction performed on 9,262 unlabeled three-dimensional CT volumes. Both the trained encoder and decoder will be used in subsequent stages. ② Training a *Diffusion Model*—a specific type of generative models—using latent features and tumor masks as conditions. Once trained, this model can generate latent features necessary for reconstructing CT volumes with tumors based on arbitrary masks. ③ Training a *Segmentation Model* using CT volumes of synthetic tumors, which are reconstructed by the decoder. With a large repository of healthy CT volumes, our DiffTumor framework can produce a vast array of synthetic tumors, varying in location, size, shape, texture, and intensity, therefore fostering high-performing AI models for tumor detection/segmentation.

where \( t \in [1, 2, \ldots, T] \) denotes the timestep and \( \beta_{1:T} \) is the variance schedule of noise.

In the inference, we synthesize the latent feature of CT volumes by sampling from \( p\left(z_{t|z_{0}}\right)\), which is approximated by recursively sampling from \( p\left(z_{t-1|z_{t}, z_{0}}\right)\). The training objective of our diffusion model [34] is as follows:

\[
E_{z_{0} \sim N(0,1), t} \left[ \left\| \epsilon - \epsilon_{\theta}\left(z_{t}, z_{0}^{\text{healthy}}, m, t\right) \right\|_{2}^{2} \right],
\]

where \( \epsilon_{\theta}(\cdot, t) \) is a 3D U-Net with interleaved self-attention layers and convolutional layers [34, 60] that predicts the noise given the input. To reduce the heavy computational cost for 3D CT volumes, we factorize the self-attention over the entire 3D data to first only attend to each 2D slide and then attend to the depth dimension, inspired by 3D video Transformers [3, 5]. This design largely reduces the computation cost of the self-attention layers in the 3D U-Net.

### 3.3. Segmentation Model

We construct a large-scale database of healthy CT volumes as a basis for our tumor synthesis method. This database includes 1,246 volumes with healthy livers, 1,901 with healthy pancreases, and 1,005 with healthy kidneys, ensuring diversity across ages, genders, nationalities, and acquisition protocols. Following Hu et al. [37], we generate realistic tumor-like shapes using ellipsoids and refine them with expert radiologist feedback for clinical plausibility (implementation details in Appendix E). By combining these generated tumor masks with the healthy CT volumes (Figure 3–Stage 3), we synthesize tumors across various domains, promoting generalizability in our model.

### 4. Experiments & Results

**Real-tumor datasets.** LiTS [6], MSD-Pancreas [2], and KiTS [33] were used for training and testing Segmentation Models on the liver, pancreas, and kidneys, respectively. We performed 5-fold cross-validation on 118 tumor CT volumes for LiTS and 120 tumor CT volumes for MSD-Pancreas and KiTS.

**Healthy CT datasets.** We collect a large repository of healthy CT volumes. Due to the computational cost and memory limitation for training, we only randomly selected 120 healthy CT volumes for the kidney and pancreas, respectively. For liver, we adopt the same healthy CT volumes as Hu et al. [37]. More details about dataset and implementation can be found in Appendix E.

#### 4.1. Visual Turing Test

We conduct the Visual Turing Test on 240 CT volumes for three organs, respectively, where 120 volumes are with real tumors, and the remaining 120 volumes are with synthesized tumors by our method. Four radiologists, with varying levels of experience ranging from junior to senior and professional, are involved in this test. The total Visual Turing Test took 144 hours (2,880 CTs). Following Hu et al. [37], each sample is inspected in a 3D view to be classified as either real or synthetic, allowing for the observation of a continuous slice sequence. The testing results are shown in Table 1. All radiologists are able to identify real tumors with a high sensitivity score (above 90%). This indicates their familiarity with the characteristics of real tumors. However, the low specificity scores (below 40%) on the three types of tumors for radiologists R1 and R3 suggest that the synthetic data strongly resembles real tumors, leading to most synthetic tumors being misidentified as real ones. As for R2...
Table 1. Visual Turing Test over three organs has been conducted with four radiologists (R1–R4). Each radiologist was provided with 240 three-dimensional CT volumes of each organ, including 120 scans with real tumors and the remaining 120 with synthetic ones. Radiologists were tasked to label each CT volume as real tumors or synthetic. A lower specificity score indicates a higher number of synthetic tumors being identified as real.

Table 2. Generalizable to different organs: comparison of generalization for early-stage tumor detection under different source organs. The scores in bold represent the best performance in each domain. DiffTumor achieves the best performance in almost all domains. Furthermore, DiffTumor serves as an effective data augmentation method for real tumors in three abdominal organs, yielding substantial improvements in all-stage tumor segmentation. Additional results for different segmentation backbones with more metrics can be found in the Appendix C.

4.2. Generalizable to Different Organs

DiffTumor can generate visually realistic tumors generalizable to a range of organs, although Diffusion Model is only trained on a specific organ tumor. In order to verify the effectiveness of our model’s generalization capacity across different organs, we conducted comparative experiments across three different abdominal organs. This involved training all Segmentation Models on tumor data from a single organ, and then applying that training to the other two organs. For the results of our method, we train DiffTumor on source organ data, then utilize healthy CT volumes to synthesize tumors in the target organ, which are used for further training of Segmentation Models. To showcase the broad applicability of our synthetic data, we compare the early-stage tumor detection capabilities across three commonly used backbones. The generalization result, shown in Table 2, suggests that it is difficult for Segmentation Models trained on real data to generalize across different organs, leading to poor performance in early-stage tumor detection. Hu et al. [37] introduces a modeling-based method, which can maintain consistent sensitivity scores for the same target, regardless of the source domain. The generalization ability of DiffTumor across organs surpasses that of most models, except in the setting that generalizing tumors from kidney to liver. Moreover, we demonstrate the strength of DiffTumor used as an augmentation method for real tumors in the same organ, as shown in Table 2. In particular, there is a notable improvement of 10.7% in the Dice Similarity Coefficient (DSC) for kidney tumors when using nnU-Net backbone. Furthermore, a decrease in the standard deviations of the DSC scores suggests that Segmentation Models become more stable. The significant improvement in DSC scores across all three organs proves that DiffTumor is an effective data augmentation method to enhance the performance of Segmentation Model.

4.3. Generalizable to Different Demographics

The ability of Segmentation Model to be generalizable to different demographics is critically important. It indicates that the model can effectively process CT scans from a diverse population, including various ages, genders, and ethnicities. To affirm the enhancement of DiffTumor for Segmentation Model to detect and segment real tumors across different individuals, we evaluate the generalization ability of Segmentation Model using a proprietary dataset at Hopkins [83]. This dataset includes various real pancreatic tumors (PDAC and Cyst) from diverse patient demographics. We utilize DiffTumor with Diffusion Model trained on MSD-Pancreas to enhance Segmentation Model. More details about the dataset and experiment setting can be found in Appendix D. Figure 4 shows that our synthetic data can yield an average improvement of 6.9% in DSC and 16.4% in sensitivity with the U-Net backbone. In particular, the
Figure 4. **Generalizable to various demographics.** Tumor detection and segmentation enhancement for individuals across various age groups and genders. DiffTumor can consistently boost tumor detection and segmentation performance by a significant margin in each patient group. Results of more segmentation backbones (e.g., nnU-Net and Swin UNETR) can be found in Appendix D.

improvement for people aged 50–60 is significant, with an enhancement of 18.9% in sensitivity and 9.1% in DSC. For both males and females, there are noticeable performance improvements for tumor detection and segmentation. These results demonstrate that our synthetic data can provide valuable assistance in clinical tumor analysis for individuals across various age groups and genders.

4.4. Advantages of DiffTumor

(1) **Reduced annotations for Diffusion Model.** The quality of synthetic data produced by a generative model is typically heavily reliant on the quantity and diversity of the paired training data used during the training phase [12, 40, 65]. We study the relationship between the number of annotated real tumors needed for the Diffusion Model and the performance of the Segmentation Model. We find that the relationship between the amount of paired training data and the quality of synthetic data isn’t always linear, as shown in Figure 5. In particular, DiffTumor only requires just one annotated tumor to train the Diffusion Model and generate synthetic tumors for the subsequent training of Segmentation Model. This contradicts the typical experience in computer vision [65], which generally requires large-scale data for training. The results indicate that for training the Diffusion Model, particularly for early tumors, we can rely on a smaller number of real tumors. This finding could have important implications for the efficiency and cost-effectiveness of training DiffTumor.

(2) **Accelerated tumor synthesis.** The speed of tumor synthesis plays a crucial role in the practical application of synthetic data. Real-time synthesis can significantly speed up the training process of Segmentation Model. The speed of generating synthetic tumors in Diffusion Model is significantly influenced by the timestep $T$. We examine the impact of the timestep on the performance of Segmentation Model. The synthetic quality using DDPM sampling [34] with different timestep is illustrated in Figure 6. As can be seen, when $T = 1$, the model collapses and fails to synthesize realistic textures for both the organ and tumor textures. Consequently, using these synthetic data to train the Segmentation Model results in poor performance. However, when $T$ is increased to more than 1, the corresponding texture can be well-generated, leading to good performance in the Segmentation Model. In consideration of the trade-off between performance and efficiency, we default to a timestep of $T = 4$ for early tumor synthesis. This balance allows for the generation of high-quality synthetic data while maintaining a reasonable efficiency level.

(3) **Improved early tumor detection.** Detecting tumors in their early stages can greatly increase the chances of successful treatment and survival. However, obtaining early-stage cancer data is challenging in practice and such cases in real datasets remain scarce. This limits the AI model’s ability to detect early tumors. As shown in Figure 7, there are several failure cases for Segmentation Model trained on real data. However, with the incorporation of our synthetic data, Segmentation Models’ capability to detect early-stage tumors improves significantly. This is one of the primary reasons why DiffTumor can achieve the best performance as displayed in Table 2. This demonstrates the value and efficacy of synthetic data in enhancing early tumor detection.
Figure 6. **Accelerated tumor synthesis.** The speed of generating synthetic tumors in Diffusion Model is significantly influenced by the timestep hyper-parameter (see figure). A faster tumor synthesis is preferred in Stage ③ when training Segmentation Model, so we examine the impact of timestep on the performance of Segmentation Model. Our findings indicate that a timestep of 4, generating a tumor in 0.2 seconds, provides the most favorable results among various tested timesteps. Considering the trade-off between performance and efficiency, we chose a timestep of 4 for this study.

5. Related Work

**Generative models** such as Energy-Based Models [19, 51, 94], Variational Autoencoders (VAE) [46–48], Generative Adversarial Networks (GAN) [10, 17, 28–30, 42, 88], and normalizing flows [49, 63, 89] have shown significant potential in creating realistic images. Among these, Diffusion Models [34, 72, 73] and their variants [45, 67, 75] have recently emerged as particularly advanced in image generation. In the medical field, generative models have been effectively utilized for tasks like image-to-image translation [58, 59, 62], reconstruction [74, 85], segmentation [23, 44, 81], image denoising [27], and anomaly detection [70, 82, 84]. In this work, we focus on generating tumors in abdominal organs based on the textures of the surrounding organs, which significantly reduces the annotated data required for training. Refer to Appendix F for a more comprehensive comparison with our DiffTumor.

**Tumor synthesis** that is widely effective for a variety of organs is an attractive topic. Successful works related to tumor synthesis based on various medical modalities include colon polyp synthesis in colonoscopy videos [69], tumor cell synthesis in fluorescence microscopy images [35], synthesized brain tumors [7, 91] and myocardial pathology [90] in MRI, lung nodule synthesis in CT images [31, 41, 86], and lesion in dermoscopic images [18]. Additionally, there are many works on synthesizing non-cancerous lesions such as COVID-19 lesion synthesis in chest CT [56, 87], and diabetic lesion synthesis in retinal images [78]. Recent studies have improved the realism of synthetic tumors in the liver [36, 38, 57, 93] and pancreas [53, 80]. AI trained on these synthetic tumors perform similarly well as those trained with real tumors. However, these methods need to be redesigned for tumors in other organs, which severely limits the generalization capabilities. In this work, we learn the tumor distribution based on generative models, i.e., Diffusion Models, to realize generalizable tumor synthesis.

6. Conclusion

This work introduces DiffTumor for generalizable tumor synthesis. We leverage the observation that early-stage tumors share similar imaging characteristics in CT scans across different organs (e.g., liver, pancreas, kidneys). As a result, DiffTumor trained solely on annotated liver tumors can directly synthesize tumors in other organs with limited annotated data (e.g., pancreas, kidney). By augmenting large-scale datasets of healthy organs (readily available in clinical settings) with these synthetic tumors, we substantially expand training data for tumor segmentation models. This augmentation significantly improves AI generalizability across diverse hospital systems and patient populations.

**Acknowledgments.** This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award. We thank Yuxiang Lai, Qian Yu, and Wenxuan Li for their constructive suggestions at several stages of the project.
References


This appendix 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

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](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

<table>
<thead>
<tr>
<th>nnU-Net [39]</th>
<th>source \ target</th>
<th>liver</th>
<th>pancreas</th>
<th>kidneys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>real tumors</td>
<td>77.4</td>
<td>1.8</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Hu et al. [37]</td>
<td>78.0</td>
<td>56.3</td>
<td>61.9</td>
</tr>
<tr>
<td></td>
<td>DiffTumor</td>
<td>80.9</td>
<td>60.7</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td>pancreas</td>
<td>1.5</td>
<td>67.0</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Hu et al. [37]</td>
<td>76.2</td>
<td>68.8</td>
<td>61.9</td>
</tr>
<tr>
<td></td>
<td>DiffTumor</td>
<td>72.8</td>
<td>75.0</td>
<td>81.0</td>
</tr>
<tr>
<td></td>
<td>kidney</td>
<td>0.9</td>
<td>0.9</td>
<td>59.5</td>
</tr>
<tr>
<td></td>
<td>Hu et al. [37]</td>
<td>76.2</td>
<td>56.3</td>
<td>69.0</td>
</tr>
<tr>
<td></td>
<td>DiffTumor</td>
<td>77.4</td>
<td>63.4</td>
<td>76.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Swin UNETR [32]</th>
<th>source \ target</th>
<th>liver</th>
<th>pancreas</th>
<th>kidneys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>real tumors</td>
<td>76.2</td>
<td>2.7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Hu et al. [37]</td>
<td>79.2</td>
<td>63.4</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td>DiffTumor</td>
<td>83.1</td>
<td>69.6</td>
<td>73.8</td>
</tr>
<tr>
<td></td>
<td>pancreas</td>
<td>1.6</td>
<td>70.5</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>Hu et al. [37]</td>
<td>66.4</td>
<td>73.2</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td>DiffTumor</td>
<td>76.2</td>
<td>79.5</td>
<td>90.5</td>
</tr>
<tr>
<td></td>
<td>kidney</td>
<td>0.6</td>
<td>0</td>
<td>69.0</td>
</tr>
<tr>
<td></td>
<td>Hu et al. [37]</td>
<td>66.4</td>
<td>63.4</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>DiffTumor</td>
<td>76.2</td>
<td>61.6</td>
<td>81.0</td>
</tr>
</tbody>
</table>

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.
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.

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

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

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

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 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). DiffTumor consistently 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.

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

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

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

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>
<thead>
<tr>
<th></th>
<th>metrics</th>
<th>fold0</th>
<th>fold1</th>
<th>fold2</th>
<th>fold3</th>
<th>fold4</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U-Net</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>real tumors</td>
<td>DSC (%)</td>
<td>75.1</td>
<td>68.0</td>
<td>69.0</td>
<td>78.1</td>
<td>70.6</td>
<td>72.0</td>
</tr>
<tr>
<td></td>
<td>NSD (%)</td>
<td>68.4</td>
<td>59.0</td>
<td>57.7</td>
<td>68.3</td>
<td>62.4</td>
<td>63.2</td>
</tr>
<tr>
<td>DiffTumor</td>
<td>DSC (%)</td>
<td>84.2</td>
<td>76.7</td>
<td>79.4</td>
<td>80.6</td>
<td>74.1</td>
<td>79.0</td>
</tr>
<tr>
<td></td>
<td>NSD (%)</td>
<td>76.6</td>
<td>64.5</td>
<td>70.7</td>
<td>71.7</td>
<td>65.8</td>
<td>69.9</td>
</tr>
<tr>
<td><strong>nnU-Net</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>real tumors</td>
<td>DSC (%)</td>
<td>73.8</td>
<td>76.8</td>
<td>80.0</td>
<td>80.5</td>
<td>73.4</td>
<td>76.9</td>
</tr>
<tr>
<td></td>
<td>NSD (%)</td>
<td>62.7</td>
<td>70.2</td>
<td>71.2</td>
<td>70.8</td>
<td>67.5</td>
<td>68.5</td>
</tr>
<tr>
<td>DiffTumor</td>
<td>DSC (%)</td>
<td>84.5</td>
<td>83.4</td>
<td>81.6</td>
<td>83.9</td>
<td>77.3</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>NSD (%)</td>
<td>78.3</td>
<td>74.4</td>
<td>74.1</td>
<td>76.9</td>
<td>72.3</td>
<td>75.2</td>
</tr>
<tr>
<td><strong>Swin UNETR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>real tumors</td>
<td>DSC (%)</td>
<td>80.6</td>
<td>64.9</td>
<td>79.1</td>
<td>76.0</td>
<td>72.2</td>
<td>74.6</td>
</tr>
<tr>
<td></td>
<td>NSD (%)</td>
<td>74.2</td>
<td>55.5</td>
<td>68.2</td>
<td>67.4</td>
<td>64.8</td>
<td>66.0</td>
</tr>
<tr>
<td>DiffTumor</td>
<td>DSC (%)</td>
<td>85.1</td>
<td>77.2</td>
<td>81.2</td>
<td>85.7</td>
<td>79.9</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>NSD (%)</td>
<td>79.2</td>
<td>70.8</td>
<td>74.1</td>
<td>78.0</td>
<td>74.1</td>
<td>75.2</td>
</tr>
</tbody>
</table>

*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 weights5 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 anaxcodes, and isotropic spacing is applied to resample each scan, resulting in a uniform voxel size of $1.0 \times 1.0 \times 1.0 \text{mm}^3$. 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 $\beta_1$ and $\beta_2$ 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 $\beta_1$ and $\beta_2$ 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 MONAI6. In this study, we implement Swin UNETR based on the Swin UNETR Base variant. The orientation of CT scans is adjusted to specific anaxcodes. Isotropic spacing is utilized to resample each scan to achieve a uniform voxel size of $1.0 \times 1.0 \times 1.0 \text{mm}^3$. 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

5https://github.com/ljwztc/CLIP-Driven-Universal-Model
6https://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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>liver DSC (%)</th>
<th>liver NSD (%)</th>
<th>pancreas DSC (%)</th>
<th>pancreas NSD (%)</th>
<th>kidneys DSC (%)</th>
<th>kidneys NSD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>real tumors</td>
<td>62.3</td>
<td>63.4</td>
<td>56.0</td>
<td>51.0</td>
<td>75.1</td>
<td>68.4</td>
</tr>
<tr>
<td>Hu et al. [37]</td>
<td>69.7</td>
<td>70.9</td>
<td>55.9</td>
<td>49.9</td>
<td>80.8</td>
<td>71.0</td>
</tr>
<tr>
<td>DiffTumor</td>
<td>70.9</td>
<td>71.2</td>
<td>64.8</td>
<td>60.5</td>
<td>84.2</td>
<td>76.6</td>
</tr>
</tbody>
</table>

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. 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.