Towards Generalizable Tumor Synthesis

Qi Chen[1] Xiaoxi Chen[2] Haorui Song[3] Zhiwei Xiong[1]
Alan Yuille[3] Chen Wei[3] Zongwei Zhou[3,*]

[1]University of Science and Technology of China
[2]Shanghai Jiao Tong University
[3]Johns 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 ($N \times M$). 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 $N \times M$ to $1 \times M$.

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-

*Correspondence to Zongwei Zhou (ZZHOU82@JH.EDU)

[1]Note 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.

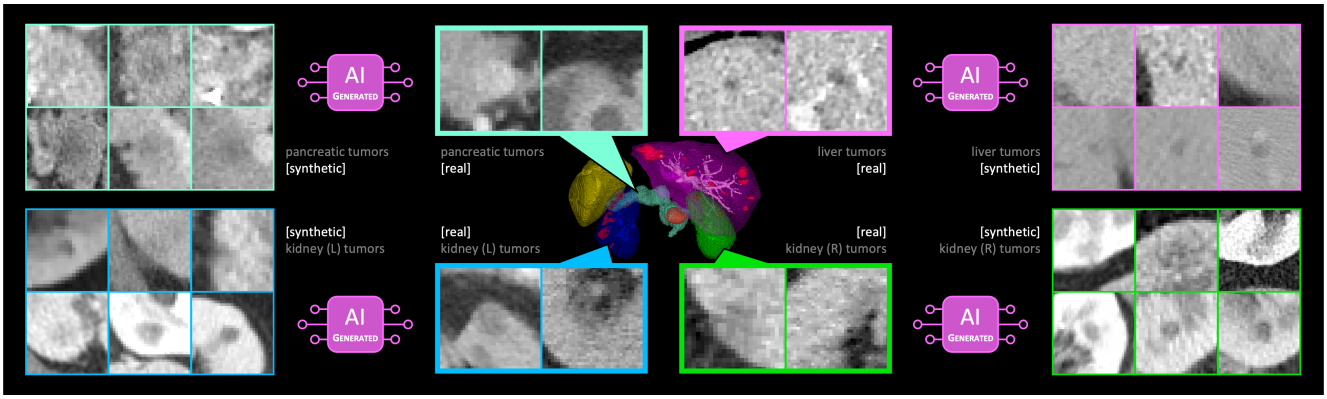


Figure 1. **Generalizable tumor synthesis across organs.** Early-stage tumors present similar imaging characteristics in computed tomography (CT), whether they are located in the liver, pancreas, or kidneys. Leveraging this observation, we develop a generative AI model on a few examples of annotated 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.

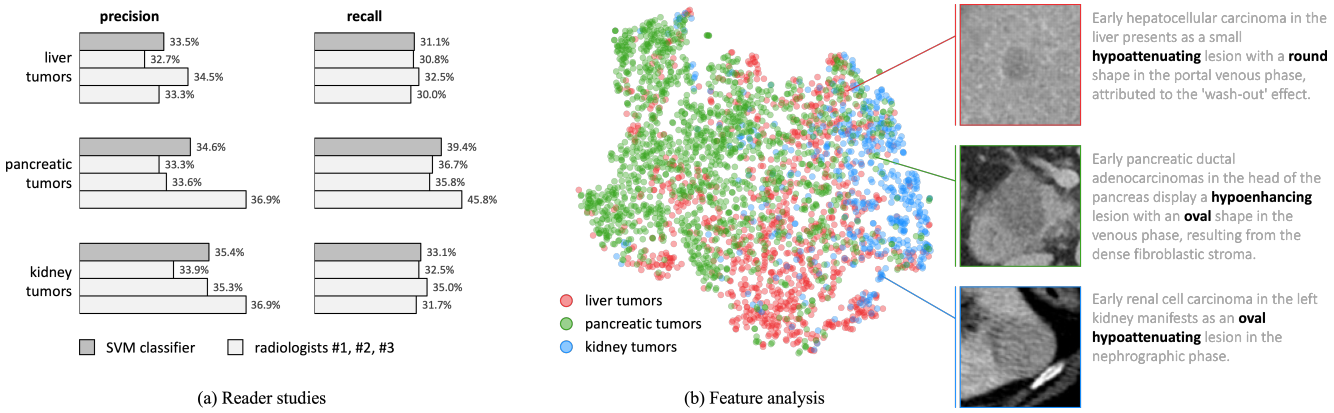
The key contributions of this paper are two-fold. **Firstly**, we have verified with feature analysis, reader studies, and clinical knowledge that early-stage tumors (<2cm) manifest with similar imaging characteristics across various organs in CT volumes, establishing the foundation for the development 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].



(a) Reader studies (b) Feature analysis

Figure 2. **Reader studies and feature analysis.** 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[3] 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

[3]We utilize the official Radiomics feature repository [76, 77] to extract the appearance features, which include 3D shape-based features, gray level co-occurrence matrix, gray level run length matrix, gray level size zone matrix, neighboring gray-tone difference matrix, and gray level dependence matrix. More details can be seen in Appendix B.

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 neoangiogenesis leads to reduced portal triads and an isoattenuating 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, hypoattenuating 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 $\boldsymbol{x} \in \mathbb{R}^{H \times W \times D}$, where H denotes the height, W the width, and D the depth. The CT sub-volume $\boldsymbol{x}$ is first converted to latent features $\boldsymbol{z}$ by an encoder f and a quantization operation $\mathbf{q}$, *i.e.*, $\boldsymbol{z} = \mathbf{q}\left(f(\boldsymbol{x})\right) \in \mathbb{Z}^{h \times w \times d}$, where h denotes the feature height, w the feature width, and d the feature depth. In the vector quantization step, the latent features $\boldsymbol{z}$ are quantized into $\boldsymbol{c}_z \in \mathbb{R}^{h \times w \times d \times c}$ by replacing each one with its closest corresponding codebook vector in the learned codebook $\mathcal{C} = \{\boldsymbol{c}_i\}_{i=1}^{K}$. K is the codebook size. Finally, a decoder g reconstructs the latent from $\boldsymbol{c}_z$ to $\hat{\boldsymbol{x}} = g\left(\boldsymbol{c}_z\right)$. The loss is a summation of three terms:

$$\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{codebook}} + \alpha \mathcal{L}_{\text{commit}}, \tag{1}$$

where $\mathcal{L}_{\text{recon}} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1$, $\mathcal{L}_{\text{codebook}} = \|\text{sg}\left[f(\boldsymbol{x})\right] - \boldsymbol{c}_z\|_2^2$, and $\mathcal{L}_{\text{commit}} = \alpha \left\|\text{sg}\left[\boldsymbol{c}_z\right] - f(\boldsymbol{x})\right\|_2^2$. $\text{sg}[\cdot]$ denotes the stop-gradient operation and α 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 $\hat{\boldsymbol{x}}$, and a 2D slice discriminator D_s to encourage per-slice quality. To stabilize the GAN training, we add the feature matching losses $\mathcal{L}_{\text{match}}$. Moreover, due to the CT volumes being preprocessed to isotropic volume, we constrain the high-frequency texture for all three planes of $\hat{\boldsymbol{x}}$ by using perceptual loss for projected reconstruction slices $\hat{\boldsymbol{x}}_{HW} \in \mathbb{R}^{H \times W}$, $\hat{\boldsymbol{x}}_{HD} \in \mathbb{R}^{H \times D}$, $\hat{\boldsymbol{x}}_{WD} \in \mathbb{R}^{W \times D}$. The overall objective of the Autoencoder is:

$$\begin{aligned} &\min_{f,g,\mathcal{C}} \left(\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{codebook}} + \alpha \mathcal{L}_{\text{commit}} + \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{perceptual}}\right) \\ &+ \min_{f,g,\mathcal{C}} \left(\max_{D_s, D_v} \left(\mathcal{L}_{\text{disc}}\right)\right), \end{aligned} \tag{2}$$

where

$$\mathcal{L}_{\text{disc}} = \log D_{s/v}(\boldsymbol{x}) + \log\left(1 - D_{s/v}(\hat{\boldsymbol{x}})\right), \tag{3}$$

$$\mathcal{L}_{\text{match}} = \sum_i \left\| D_{s/v}^{(i)}(\hat{\boldsymbol{x}}) - D_{s/v}^{(i)}(\boldsymbol{x}) \right\|_1. \tag{4}$$

$D_{s/v}^{(i)}$ 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 $\boldsymbol{x}_0$ and the mask of its tumor region $\boldsymbol{m}$, the diffusion model is conditioned on both the tumor mask $\boldsymbol{m}$ and the healthy region $\boldsymbol{z}_0^{\text{healthy}} := (1 - \boldsymbol{m}) \odot \boldsymbol{z}_0$. The diffusion model approximates the distribution of the latent features of tumor-present CT volumes. In the forward process, the latent feature $\boldsymbol{z}_0$ is gradually converted to white Gaussian noise $\boldsymbol{z}_T \sim \mathcal{N}(0, 1)$ by recursively adding a small amount of Gaussian noise T times following the Markov process below:

$$p\left(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{z}_t; \sqrt{1 - \beta_t}\boldsymbol{z}_{t-1}, \beta_t \mathbf{I}\right), \tag{5}$$

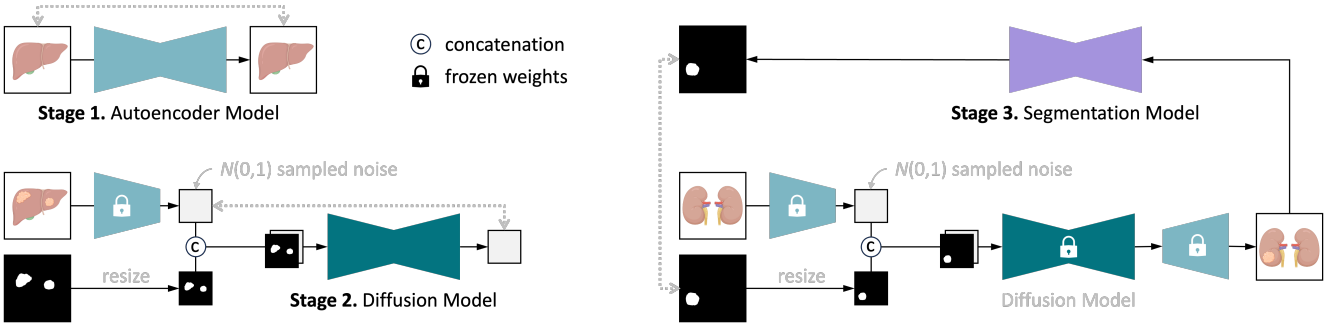


Figure 3. **Overview of the DiffTumor framework.** Towards generalizable tumor synthesis, developing our DiffTumor involves three stages. ① Training an *Antoencoder Model*—consisting of an encoder and decoder—to learn comprehensive latent features. The learning task here is image reconstruction performed 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 $t \in [1, 2, .., T]$ 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 $p\left(\boldsymbol{z}_0 \mid \boldsymbol{z}_0^{\text{healthy}}, \boldsymbol{m}\right)$, which is approximated by recursively sampling from $p\left(\boldsymbol{z}_{t-1}, \mid \boldsymbol{z}_t, \boldsymbol{z}_0^{\text{healthy}}, \boldsymbol{m}\right)$. The training objective of our diffusion model [34] is as follows:

$$\mathbb{E}_{\boldsymbol{z}_0, \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta\left(\boldsymbol{z}_t, \boldsymbol{z}_0^{\text{healthy}}, \boldsymbol{m}, t\right)\right\|_2^2\right], \tag{6}$$

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

		R1	R2	R3	R4
liver	sensitivity (%)	98.3	99.2	100	100
	specificity (%)	31.7	53.3	39.2	45.8
	accuracy (%)	65.0	76.3	69.6	72.9
pancreas	sensitivity (%)	96.7	100	100	98.3
	specificity (%)	22.5	44.2	34.2	38.8
	accuracy (%)	59.6	72.1	67.1	68.3
kidney	sensitivity (%)	95.8	98.3	99.2	97.5
	specificity (%)	36.7	55.0	40.8	51.7
	accuracy (%)	66.3	76.7	70.0	74.6

positives: real tumors (N = 120); negatives: synthetic tumors (N = 120).

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* or *synthetic*. A lower specificity score indicates a higher number of synthetic tumors being identified as real.

and R4, who have more experience, the specificity scores are higher than those of R1 and R3, approximating 50%. This indicates that nearly 50% of synthetic samples are still incorrectly identified as real tumors. These results confirm the efficacy of DiffTumor in generating visually realistic tumors.

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 method'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

Early-stage tumor detection performance (tumor-wise Sensitivity %).

source \ target		liver	pancreas	kidneys
liver	real tumors	75.6	0	2.4
	Hu *et al*. [37]	77.8	56.3	52.4
	DiffTumor	**82.2**	**56.3**	**76.2**
pancreas	real tumors	0.7	64.3	0
	Hu *et al*. [37]	74.1	67.0	52.4
	DiffTumor	**75.3**	**71.4**	**71.4**
kidney	real tumors	0.1	0	50.0
	Hu *et al*. [37]	**74.1**	56.3	66.7
	DiffTumor	68.8	**61.6**	**78.6**

All-stage tumor segmentation performance (DSC %).

backbone	method	liver	pancreas	kidneys
U-Net	real tumors	62.3±28.3	56.0±24.8	75.1±27.2
	DiffTumor	**70.9±21.1**	**64.8±24.5**	**84.2±9.5**
nnU-Net	real tumors	64.3±26.5	59.9±23.8	73.8±20.9
	DiffTumor	**73.6±18.1**	**63.6±27.7**	**84.5±11.5**
SwinUNETR	real tumors	65.1±23.5	52.2±31.2	80.6±19.6
	DiffTumor	**71.4±19.1**	**62.2±26.1**	**85.1±8.7**

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.

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

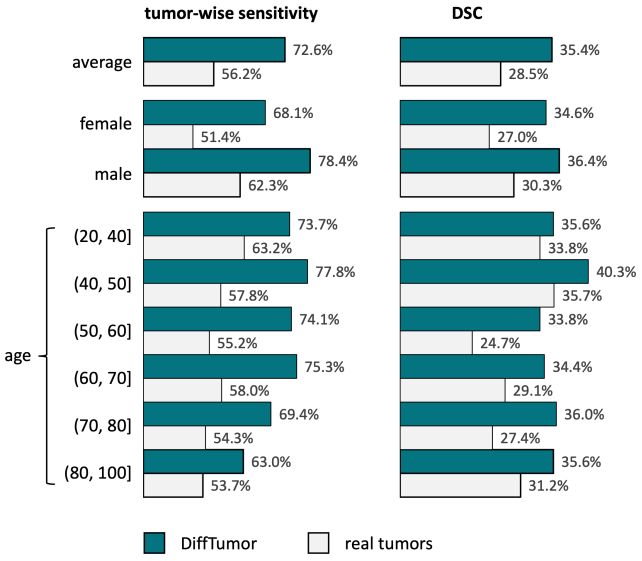


Figure 4. **Generalizable to various demographics.** Tumor detection and segmentation enhancement for individuals across various age groups and genders. DiffTumor can consistently boost tumor detection and segmentation performance by a significant margin in each patient group. Results of more segmentation 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.

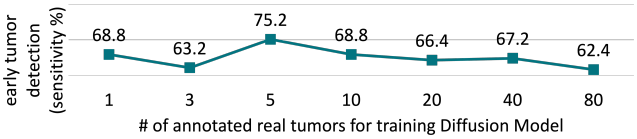


Figure 5. **Reduced annotations for Diffusion Model.** Diffusion Model, trained on annotated tumors in Stage ②, can generate synthetic tumors for the subsequent training of Segmentation Model in Stage ③. We investigate the relationship between the number of annotated real tumors required for Diffusion Model and the resultant performance of Segmentation Model. Results with varying numbers of annotated tumors reveal a surprising finding: extensive annotations are not necessary for tumor synthesis, contrary to the experience in computer vision [67, 68]. Notably, training Diffusion Model with only one annotated tumor seems to be sufficient. This efficiency is connected to our earlier observation in §2 that tumors, particularly in their early stages, tend to present similar appearances across different organs, thus facilitating the learning process of Diffusion Model with fewer annotated examples.

(2) Accelerated tumor synthesis. The speed of tumor synthesis plays a crucial role in the practical application of synthetic data. Real-time synthes 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.

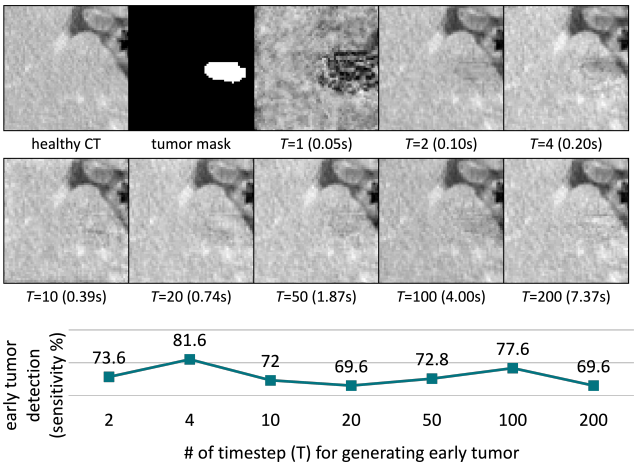


Figure 6. **Accelerated tumor synthesis.** The speed of generating synthetic tumors in Diffusion Model is significantly influenced by the timestep hyper-parameter (see figure). A faster tumor synthesis is preferred in Stage ③ when training Segmentation Model, so we examine the impact of timestep on the 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.

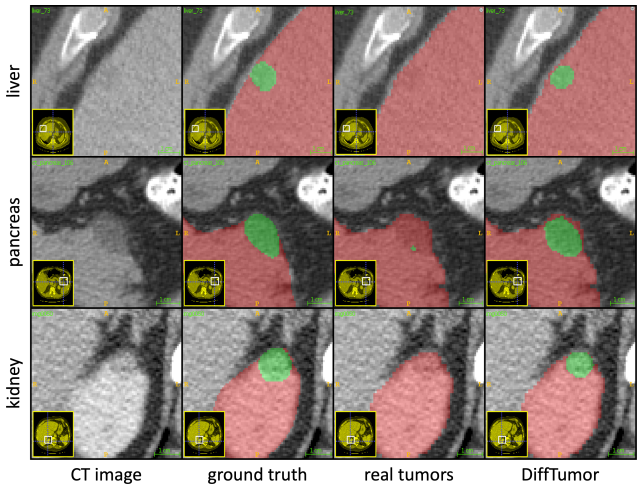


Figure 7. **Enhanced early tumor detection.** We analyzed failure cases of AI models trained on real tumors, identifying that these models often overlook early-stage tumors characterized by blurry boundaries, small sizes, and peripheral organ locations. Our DiffTumor enhances the detection and segmentation of these challenging tumors by extensively generating small tumors for AI training (evidenced in Table 2). Further visualizations and examples are available in Appendix C.

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 dermatoscopic 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

[1] Clément Abi Nader, Rebeca Vetil, Laura Kate Wood, Marc-Michel Rohe, Alexandre Bône, Hedvig Karteszi, and Marie-Pierre Vullierme. Automatic detection of pancreatic lesions and main pancreatic duct dilatation on portal venous ct scans using deep learning. *Investigative Radiology*, 2023. 1

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. 5, 21

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 5

[4] Carmen Ayuso, Jordi Rimola, Ramón Vilana, Marta Burrel, Anna Darnell, Ángeles García-Criado, Luis Bianchi, Ernest Belmonte, Carla Caparroz, Marta Barrufet, et al. Diagnosis and staging of hepatocellular carcinoma (hcc): current guidelines. *European journal of radiology*, 101:72–81, 2018. 4

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 5

[6] 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. 5, 21

[7] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023. 8

[8] Harry B Burke. Outcome prediction and the future of the tnm staging system, 2004. 2

[9] Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, et al. Cancerunit: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21327–21338, 2023. 1

[10] Qi Chen, Mingxing Li, Jiacheng Li, Bo Hu, and Zhiwei Xiong. Mask rearranging data augmentation for 3d mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2022. 8

[11] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. 1

[12] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021. 7

[13] Jin-Young Choi, Jeong-Min Lee, and Claude B Sirlin. Ct and mr imaging diagnosis and staging of hepatocellular carcinoma: part i. development, growth, and spread: key pathologic and imaging aspects. *Radiology*, 272(3):635–654, 2014. 2, 4

[14] Yu-Cheng Chou, Bowen Li, Deng-Ping Fan, Alan Yuille, and Zongwei Zhou. Acquiring weak annotations for tumor localization in temporal and volumetric data. *Machine Intelligence Research*, pages 1–13, 2024. 1

[15] Linda C Chu, Michael G Goggins, and Elliot K Fishman. Diagnosis and detection of pancreatic cancer. *The Cancer Journal*, 23(6):333–342, 2017. 4

[16] Linda C Chu, Seyoun Park, Satomi Kawamoto, Daniel F Fouladi, Shahab Shayesteh, Eva S Zinreich, Jefferson S Graves, Karen M Horton, Ralph H Hruban, Alan L Yuille, et al. Utility of ct radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue. *American Journal of Roentgenology*, 213(2):349–357, 2019. 3

[17] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 8

[18] Shiyi Du, Xiaosong Wang, Yongyi Lu, Yuyin Zhou, Shaoting Zhang, Alan Yuille, Kang Li, and Zongwei Zhou. Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. *arXiv preprint arXiv:2310.02906*, 2023. 8

[19] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[20] N Reed Dunnick. Renal cell carcinoma: staging and surveillance. *Abdominal Radiology*, 41:1079–1085, 2016. 4

[21] Khaled Y Elbanna, Hyun-Jung Jang, and Tae Kyoung Kim. Imaging diagnosis and staging of pancreatic ductal adenocarcinoma: a comprehensive review. *Insights into imaging*, 11(1):1–13, 2020. 4

[22] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[23] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosiu, Mark S Graham, Tom Vercauteren, and M Jorge Cardoso. Can segmentation models be trained with fully synthetically generated data? In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 79–90. Springer, 2022. 8

[24] Vincenzo Ficarra, Antonio Galfano, Mariangela Mancini, Guido Martignoni, and Walter Artibani. Tnm staging system for renal-cell carcinoma: current status and future perspectives. *The lancet oncology*, 8(6):554–558, 2007. 2

[25] Kathryn J Fowler, Adam Burgoyne, Tyler J Fraum, Mojgan Hosseini, Shintaro Ichikawa, Sooah Kim, Azusa Kitao, Jeong Min Lee, Valérie Paradis, Bachir Taouli, et al.

Pathologic, molecular, and prognostic radiologic features of hepatocellular carcinoma. *Radiographics*, 41(6):1611–1631, 2021. 4

[26] Cong Gao, Benjamin D Killeen, Yicheng Hu, Robert B Grupp, Russell H Taylor, Mehran Armand, and Mathias Unberath. Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308, 2023. 1

[27] Kuang Gong, Keith Johnson, Georges El Fakhri, Quanzheng Li, and Tinsu Pan. Pet image denoising based on denoising diffusion probabilistic model. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–11, 2023. 8

[28] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 8

[29] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 8

[31] 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. 8

[32] 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, 2021. 16

[33] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020. 5, 21

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 5, 7, 8

[35] 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. 8

[36] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jieneng Chen, Alan Yuille, and Zongwei Zhou. Synthetic tumors make ai segment tumors better. *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. 8

[37] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023. 5, 6, 16, 22

[38] Qixin Hu, Alan Yuille, and Zongwei Zhou. Synthetic data as validation. *arXiv preprint arXiv:2310.16052*, 2023. 8

[39] 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. 16

[40] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020. 7

[41] 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. 8

[42] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018. 1, 8

[43] Mintong Kang, Bowen Li, Zengle Zhu, Yongyi Lu, Elliot K Fishman, Alan Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 1

[44] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. *arXiv preprint arXiv:2209.14566*, 2022. 8

[45] Dongjun Kim, Byeonghu Na, Se Jung Kwon, Dongsoo Lee, Wanmo Kang, and Il-chul Moon. Maximum likelihood training of parametrized diffusion model. 2021. 8

[46] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 8

[47] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[48] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 8

[49] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020. 8

[50] Paul F Laeseke, Ru Chen, R Brooke Jeffrey, Teresa A Brentnall, and Jürgen K Willmann. Combining in vitro diagnostics with in vivo imaging for earlier detection of pancreatic ductal adenocarcinoma: challenges and solutions. *Radiology*, 277 (3):644–661, 2015. 4

[51] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 8

[52] Michael J Leveridge, Peter J Bostrom, George Koulouris, Antonio Finelli, and Nathan Lawrentschuk. Imaging renal cell carcinoma with ultrasonography, ct and mri. *Nature Reviews Urology*, 7(6):311–325, 2010. 4

[53] Bowen Li, Yu-Cheng Chou, Shuwen Sun, Hualin Qiao, Alan Yuille, and Zongwei Zhou. Early detection and localization of pancreatic cancer by label-free tumor synthesis. *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023. 8

[54] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *The Twelfth International Conference on Learning Representations*, 2024. 1

[55] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 1, 21, 22

[56] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. *IEEE Transactions on Medical Imaging*, 2022. 1, 8

[57] Fei Lyu, Mang Ye, Andy J Ma, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and Pong C Yuen. Learning from synthetic ct images via test-time training for liver tumor segmentation. *IEEE transactions on medical imaging*, 41(9):2510–2520, 2022. 8

[58] Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022. 8

[59] Xiangxi Meng, Yuning Gu, Yongsheng Pan, Nizhuan Wang, Peng Xue, Mengkang Lu, Xuming He, Yiqiang Zhan, and Dinggang Shen. A novel unified conditional score-based generative framework for multi-modal medical image completion. *arXiv preprint arXiv:2207.03430*, 2022. 8

[60] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 5

[61] Mauricio Orbes-Arteaga, Thomas Varsavsky, Carole H Sudre, Zach Eaton-Rosen, Lewis J Haddow, Lauge Sørensen, Mads Nielsen, Akshay Pai, Sébastien Ourselin, Marc Modat, et al. Multi-domain adaptation in brain mri through paired consistency and adversarial learning. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 54–62. Springer, 2019. 1

[62] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 8

[63] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021. 8

[64] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. *Conference on Neural Information Processing Systems*, 2023. 21

[65] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 7

[66] Guido Rindi, Massimo Falconi, Catherine Klersy, L Albarello, L Boninsegna, MW Buchler, C Capella, Martyn Caplin, Anne Couvelard, Claudio Doglioni, et al. Tnm staging of neoplasms of the endocrine pancreas: results from a large international cohort study. *Journal of the National Cancer Institute*, 104(10):764–777, 2012. 2

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 7, 8

[68] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 7

[69] Younghak Shin, Hemin Ali Qadir, and Ilangko Balasingham. Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. *IEEE Access*, 6:56007–56017, 2018. 1, 8

[70] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 191–200, 2019. 8

[71] Arthur T Skarin. *Atlas of Diagnostic Oncology E-Book*. Elsevier Health Sciences, 2015. 4

[72] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 8

[73] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 8

[74] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 8

[75] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 8

[76] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21): e104–e107, 2017. 3, 15

[77] Hongkai Wang, Zongwei Zhou, Yingci Li, Zhonghua Chen, Peiou Lu, Wenzhi Wang, Wanyu Liu, and Lijuan Yu. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18f-fdg pet/ct images. *EJNMMI research*, 7(1):1–11, 2017. 3

[78] 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*, page 102534, 2022. 1, 8

[79] Z. J. Wang, A. C. Westphalen, and R. J. Zagoria. CT and MRI of small renal masses. *Br J Radiol*, 91(1087):20180131, 2018. 4

[80] Zihan Wei, Yizhou Chen, Qiu Guan, Haigen Hu, Qianwei Zhou, Zhicheng Li, Xinli Xu, Alejandro Frangi, and Feng Chen. Pancreatic image augmentation based on local region texture synthesis for tumor segmentation. In *International Conference on Artificial Neural Networks*, pages 419–431. Springer, 2022. 8

[81] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022. 8

[82] 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. 8

[83] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022. 1, 6

[84] Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23890–23901, 2023. 8

[85] Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer, 2022. 8

[86] Jie Yang, Siqi Liu, Sasa Grbic, Arnaud Arindra Adiyoso Setio, Zhoubing Xu, Eli Gibson, Guillaume Chabin, Bogdan Georgescu, Andrew F Laine, and Dorin Comaniciu. Class-aware adversarial lung nodule synthesis in ct images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1348–1352. IEEE, 2019. 8

[87] 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*, 2021. 1, 8

[88] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019. 1, 8

[89] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 8

[90] Hantao Zhang, Jiancheng Yang, Shouhong Wan, and Pascal Fua. Lefusion: Synthesizing myocardial pathology on cardiac mri via lesion-focus diffusion models. *arXiv preprint arXiv:2403.14066*, 2024. 8

[91] Xiaoman Zhang, Weidi Xie, Chaoqin Huang, Ya Zhang, Xin Chen, Qi Tian, and Yanfeng Wang. Self-supervised tumor segmentation with sim2real adaptation. *IEEE Journal of Biomedical and Health Informatics*, 2023. 8

[92] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L Yuille, Yaoyao Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2023. 21

[93] Zhaoxiang Zhang, Hanqiu Deng, and Xingyu Li. Unsupervised liver tumor segmentation with pseudo anomaly synthesis. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 86–96. Springer, 2023. 8

[94] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 8

[95] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021. 1

[96] Zongwei Zhou, Michael B Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health*, pages 343–371. Springer, 2022. 1

[97] Zengle Zhu, Mintong Kang, Alan Yuille, and Zongwei Zhou. Assembling and exploiting large-scale existing labels of common thorax diseases for improved covid-19 classification using chest radiographs. In *Radiological Society of North America (RSNA)*, 2022. 1