Text-Driven Tumor Synthesis

Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Tianyu Lin, Pengfei Guo, Dong Yang, Can Zhao, Pedro R. A. S. Bassi, Daguang Xu, Kang Wang, Yang Yang, Alan L. Yuille, *Fellow, IEEE*, Zongwei Zhou, *Member, IEEE*

***Abstract*— Tumor synthesis can generate challenging cases that AI often misses or over-detects. Training on these cases improves AI performance. However, most existing synthesis methods are either unconditional—generating images from random variables—or conditioned only on tumor shape. As a result, they lack control over clinically important tumor characteristics, such as texture, heterogeneity, boundary, and pathology. The generated tumors are therefore overly similar or duplicates of existing training cases, failing to effectively address AI's weaknesses. We propose a new text-driven tumor synthesis approach, termed TextoMorph, that provides textual control over tumor characteristics in conjunction with mask control. This approach is particularly beneficial for examples that confuse the AI the most, such as early tumor detection (improving Sensitivity by +6.5%), tumor segmentation for precise radiotherapy (improving NSD by +3.1%), and classification between benign and malignant tumors (improving Sensitivity by +8.2%). By incorporating text mined from radiology reports into the synthesis process, we increase the variability and controllability of the synthetic tumors to target AI's failure cases more precisely. Moreover, TextoMorph uses contrastive learning across different texts and CT scans, significantly reducing dependence on scarce image-report pairs (only 141 pairs used in this study) by leveraging a large corpus of 34,035 radiology reports. Finally, we have developed rigorous tests to evaluate synthetic tumors, showing that our synthetic tumors is realistic and diverse in texture, heterogeneity, boundary, and pathology. Code and models are available at https://github.com/MrGiovanni/TextoMorph**

***Index Terms*— Tumor Synthesis, Generative Models, Report Mining, Tumor Detection, Tumor Classification**

I. Introduction

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. Corresponding author: Zongwei Zhou

X. Li is with the Department of Biomedical Engineering, Yale University, New Haven, CT 06520 USA (e-mail: alena.li@yale.edu).

Y. Shuai is with the The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, Guangdong 510080, China (email: shuaiy3@mail2.sysu.edu.cn).

C. Liu is with the School of Nursing, Hong Kong Polytechnic University, Hong Kong (email: marie-chen.liu@polyu.edu.hk).

T. Lin is with the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (email: tlin67@jh.edu).

P. Guo, D. Yang, C. Zhao, and D. Xu are with NVIDIA, Santa Clara, CA 95050, USA (email: pengfeig@nvidia.com; dongy@nvidia.com; canz@nvidia.com; daguangx@nvidia.com).

K. Wang and Y. Yang are with Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, 94143 USA (email: kang.wang@ucsf.edu; yang.yang4@ucsf.edu).

Q. Chen, P. Bassi, A. Yuille, and Z. Zhou are with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (email: qchen76@jh.edu; psalvad2@jh.edu; zzhou82@jh.edu; ayuille1@jhu.edu).

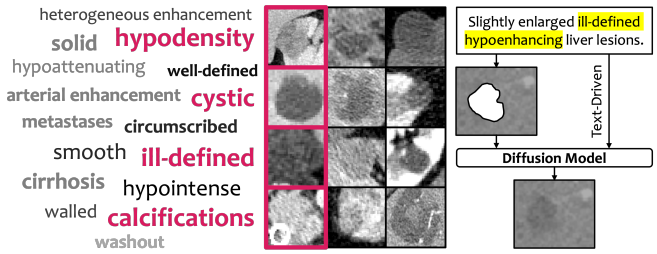


Fig. 1. **Text-Driven Tumor Synthesis.** Existing synthesis methods lack controllability and often generate tumors based only on shapes or random noise. This results in synthetic tumors that lack essential features like texture, boundaries, and attenuation, reducing its effectiveness in analyzing and addressing AI weaknesses. TextoMorph addresses this limitation by exploiting a dataset of 34,035 radiology reports to generate tumors with medically precise features described in clinical language. Examples include phrases such as '*hypodensity*', '*ill-defined*', and '*cystic*', paired with CT scans of the liver, pancreas, and kidney.

TUMOR synthesis plays a critical role in *targeted* data augmentation by generating examples that AI models tend to miss (false negatives) or over-detect (false positives), focusing on areas needing improvement [1], [2] and addressing privacy concerns and reducing annotation costs [3]–[7]. However, existing synthesis methods are typically unconditional [8]—generating images from random variables—or conditioned only on shape masks [9], lacking controls over specific tumor characteristics such as texture, heterogeneity, boundaries, and pathology type. We find that text should be considered an important conditioning factor when generating tumors because it carries much richer information[1] than random variables or shape masks can offer. Moreover, tumor-related text is readily available in radiology reports, which are routinely generated by radiologists in clinical workflows [10]–[14]. We hypothesize that incorporating text as a condition, alongside tumor masks, allows us to develop stronger AI models for tumor detection, segmentation, and classification due to greater controllability of generating such tumors that AI models often make mistakes.

Text-driven generative models [15]–[19] have significantly advanced in recent years. These models leverage natural language descriptions to control the synthesis of images/videos, enabling fine-grained manipulation of generated content. Applications range from data augmentation for AI training to commercial products that generate images/videos for creative and practical purposes [20]–[23]. However, these models have

[1]For example, a report goes '*slightly enlarged ill-defined liver lesions*' and '*more well-defined appearance of liver lesions*.' The corresponding CT scans are shown in Figure 1.

not been fully explored in tumor synthesis due to several challenges: ***First**, lack of annotated tumors:* Only a very small proportion (less than 5%) of publicly available abdominal CT datasets contain voxel-wise annotated tumors [24]–[30]. ***Second**, lack of text descriptions:* Very few publicly available datasets have radiology reports or text descriptions for tumors in CT scans [31]–[36]. ***Third**, need of large-scale paired datasets for training:* For example, DALL·E was trained on 250 million image-text pairs [37], and Imagen Video was trained on 14 million video-text pairs along with 60 million image-text pairs [38]. ***Forth**, difficulty in evaluating generated synthetic tumors:* AI-generated images/videos can be assessed by anybody, while generated tumors must be visually inspected by busy, costly medical professionals [39]–[43].

To address these challenges, we first create a dataset consisting of 141 CT-Report pairs containing tumors in the liver, pancreas, and kidney, along with 34,035 radiology reports that provide textual descriptions of tumors or normal findings (see examples in Figure 1). We then develop a new text-driven tumor synthesis approach, termed **TextoMorph**, which can generate targeted tumors based on the described tumor characteristics. By incorporating textual descriptions mined from radiology reports into the synthesis process, TextoMorph increases the variability and controllability of the synthetic tumors, allowing us to precisely target the AI's failure modes. Our TextoMorph outperforms a total of four existing generative methods. It is particularly beneficial for such cases that challenge AI the most, including (1) early-stage tumor detection (less than 20mm), increasing Sensitivity by **+6.5%** (Figure 5), (2) tumor segmentation for precise radiotherapy, increasing NSD by **+3.1%** (Figure 5), and (3) classification between benign and malignant tumors, improving Sensitivity by **+8.2%** (Table IV).

More importantly, we have also developed rigorous tests to evaluate the effectiveness of synthetic tumors for targeted data augmentation. ***First**, Text-Driven Visual Turing Test to examine tumor fidelity.* Radiologists were asked to distinguish real and synthetic tumors with the same shape mask and text description (e.g., both being *'cystic tumors'*). As shown in Table I, they erred 22.5–45.0% of the time, significantly higher than previous rates of 7.5–30.0%, suggesting that TextoMorph generates highly realistic, text-accurate synthetic tumors. ***Second**, Radiomics Pattern Analysis to analyze the diversity of generated tumor appearance.* We compute texture-wise radiomics features of synthetic tumors conditioned on different random noise. TextoMorph showed greater variance than prior arts (e.g., 1.03 for TextoMorph vs. 0.93 for DiffTumor [9]; Table II). This indicates that TextoMorph generates diverse, text-aligned, realistic tumors, explaining the robust performance of AI trained on them. These promising results are attributable to the following novel design of TextoMorph:

1) **Tumor Report Preprocessing (§III-A).** To address the lack of paired text-image data, we applied Large Language Models (LLMs) [44], [45] to automatically generate detailed descriptions highlighting tumor texture, margins, and pathology. These LLM-generated texts undergo semantic validation, ensuring consistency with original reports and providing reliable textual conditions for tumor synthesis.
2) **Text-Driven 3D Diffusion Model (§III-B).** We develop a 3D diffusion model conditioned on radiology report text to control tumor appearance, including texture, margins, and pathology. To strengthen text–image alignment, we apply contrastive learning [46], [47]: CTs paired with identical descriptions form *positive pairs* across different scans, while the same scan with differing descriptions forms *negative pairs*. The contrastive objective is applied to tumor-specific features and scaled with a corpus of 34,035 reports, yielding robust associations between textual phrases and visual tumor characteristics.
3) **Targeted Data Augmentation (§III-C).** We analyzed the model failures (e.g., false negatives) and generated tailored synthetic tumors specifically designed to address these shortcomings. Incorporating this targeted data augmentation led to notable improvements in tumor detection, segmentation, and classification performance.

II. Related Work

Tumor synthesis has emerged as a critical research focus across various medical imaging modalities, including colonoscopy videos [48], MRI [49], CT [50]–[52], and endoscopic images [53]–[55]. While early methods [50]–[52], [56]–[58] relied on low-level image processing, their limited realism often produced noisy synthetic data that degraded downstream performance. To address this, condition-guided synthesis has gained traction, enabling precise tumor localization and morphology control for stronger augmentation in detection and segmentation [59]–[64]; we include representative methods as baselines. In parallel, advances such as latent diffusion [65] and ControlNet-style conditioning [66] have substantially improved fidelity and structural controllability in general imaging. Recent medical adaptations, including MedCLIP [67], VoCo [68], and Unified Contrastive Learning [69], further explore image–text alignment and contrastive objectives for representation learning and global-level synthesis. Our framework builds upon these advances but focuses on lesion-level, mask and text conditioned generation in 3D CT to achieve descriptor-controllable tumor synthesis. Nevertheless, most medical tumor synthesis remains mask-only, offering limited control over texture, heterogeneity, and boundaries.

Text-driven synthesis has emerged as a transformative tool in medical imaging, enabling the generation of diverse medical images, such as chest X-rays, histopathology, and retinal images, based on descriptive text [70]. This approach has significantly advanced tasks like multi-abnormality classification and rare condition research, while also improving data curation efficiency through automated labeling and synthetic data generation [40], [71]–[73]. Additionally, applications in privacy-preserving analytics and digital technology further highlight its potential [74]–[76]. However, existing methods primarily focus on whole-CT-level synthesis, limiting their utility for pathology-specific tasks. To address this, we develop a novel text-driven tumor synthesis framework termed TextoMorph that enables the precise generation of tumors based on described characteristics.

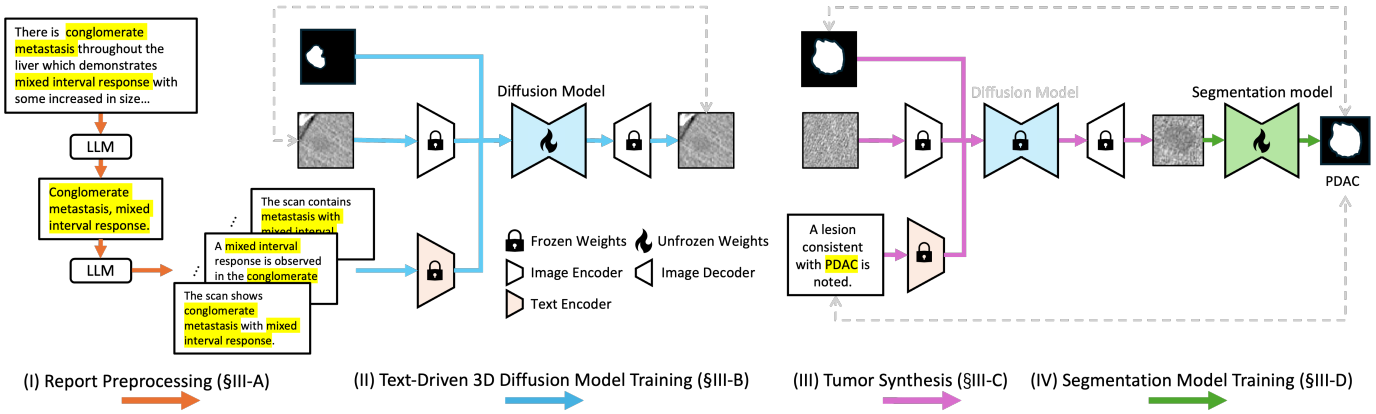


Fig. 2. **Overview of the TextoMorph Framework.** The framework consists of four steps: **(I) Tumor Report Preprocessing** (§III-A): From each radiology report, we extract descriptive phrases (highlighted in yellow) and generate radiology sentences containing these keywords. A text encoder converts these text inputs into language embeddings. **(II) Text-driven 3D Diffusion Model Training** (§III-B): A diffusion model with an image encoder and decoder is trained to synthesize high-fidelity tumors conditioned on mask, text representation, and latent CT features. **(III) Tumor Synthesis** (§III-C): Using the frozen diffusion model, we synthesize tumors jointly conditioned on large sets of masks and text descriptions (e.g., PDAC, a type of pancreatic lesion). **(IV) Segmentation Model Training** (§III-D): Large-scale synthesis of tumors is used to augment training for the segmentation model, improving downstream performance including segmentation, detection, and classification.

III. TextoMorph

A. Tumor Report Preprocessing

Radiology reports often contain fragmented and inconsistent descriptions, which makes text-controlled tumor synthesis challenging. To address this issue, we adopt a two-stage preprocessing strategy that extracts and augments tumor-specific textual descriptions.

Text Extraction. We employed GPT-4o [44] to extract tumor characteristics, focusing on texture and margins. Using prompts such as '*Extract detailed texture and margin characteristics central to the tumor field,*' we generated a cleaned descriptive output S_i. To verify accuracy, we computed cosine similarity between S_i and the original report by Llama 3.1, excluding cases with similarity below 0.9 to ensure consistency.

Text Generation. For each S_i, we generated $N = 100$ variant reports $\mathcal{R}i1, \mathcal{R}_{i2}, \ldots, \mathcal{R}_{i100}$ by varying sentence structures while carefully retaining core descriptive features. We prompted GPT-4o with '*generate 100 reports with distinct sentence structures, ensuring that critical texture and margin information is retained accurately.*' Llama 3.1 evaluated cosine similarity, excluding variants with similarity below 0.9 to rigorously maintain alignment. This systematic process expanded each CT image's textual association from one report to 100 semantically consistent variants, forming a robust dataset $\mathcal{X} = (x_i, \mathcal{R}_{ij})$ for controlled tumor synthesis enriched with diverse textual descriptions. Here, x_i denotes the i-th CT image. We selected 100 text variants to generate each augmented report $\mathcal{R}_i$. This selection is motivated by three important considerations: (i) to enhance model robustness to text variability during inference; (ii) to sufficiently account for the exponential growth in combinations of medical terms, synonyms, and anatomical substructures (e.g., *liver*, *hepatic*, *liver sub-segment*); and (iii) to effectively capture diverse real-world reporting styles across hospitals. Our experiments also confirmed that generating an excessive number of text variants risks redundancy and reduced effectiveness.

B. Text-Driven 3D Diffusion Model Training

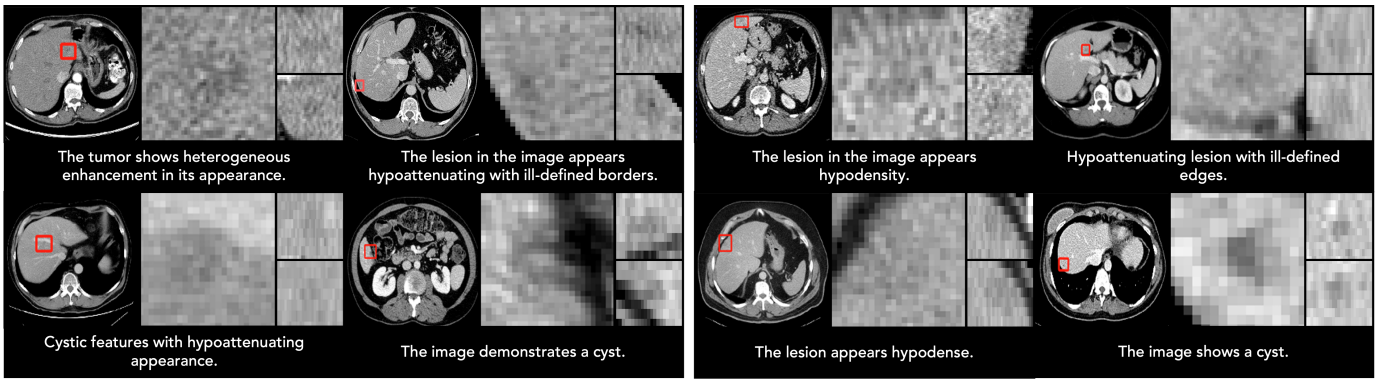
Text-Driven Tumor Generator. We adopt latent diffusion models (LDMs) [65], [77]–[79] to extract compact features from 3D CT volumes and enable controlled tumor synthesis. Each CT sub-volume $x \in \mathbb{R}^{H\times W\times D}$ is encoded by a 3D VQGAN [80] into a latent representation $z_0 = E(x)$, which is reconstructed by the decoder D to preserve important details.

Following DiffTumor [9], we apply a 200-step diffusion process that gradually corrupts z_0 with noise. A time-conditional 3D U-Net ϵ_θ learns to reverse this process. We denote the overall text-driven 3D diffusion generator as $g_\theta(\mathcal{R}, x, m)$, which synthesizes a tumor-containing CT volume conditioned on the input report $\mathcal{R}$, healthy CT x, and binary mask m. Its conditioning inputs include the healthy tissue latent $z_{\text{healthy}} = E((1-m)\odot x)$, a binary tumor mask m, and text embeddings $\tau_\theta(\mathcal{R}_i)$ derived from augmented radiology reports. We selected 100 text variants to generate each augmented report $\mathcal{R}_i$, with the rationale detailed in §III-A. The denoising network predicts the noise according to $\hat{\epsilon} = \epsilon_\theta(z_t, t, z_{\text{healthy}}, \tau_\theta(\mathcal{R}_i), m)$, and recovers the latent representation by

$$\hat{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}\right), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denotes the cumulative noise attenuation factor. The forward process is defined as $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. After 200 denoising iterations, the refined latent $\hat{z}_0$ is decoded by D to produce a CT with a tumor exhibiting the characteristics.

Text-Driven Contrastive Learning. We propose a robust text-driven contrastive loss to improve both alignment with textual descriptions and enhanced diversity of generated tumors. Given a detailed report $\mathcal{R}_i$, CT scan x_i, and corresponding mask m_i, we generate a synthetic tumor $\mathcal{T}_i = g_\theta(\mathcal{R}_i, x_i, m_i)$. To encourage greater diversity and variability, we sample a different report $\mathcal{R}_j$ and generate $\mathcal{T}_j = g_\theta(\mathcal{R}_j, x_i, m_i)$ on the same CT context, strictly maximizing their feature distance; to



(a) False negatives missed by the AI **(b) Synthetic tumors generated based on false negatives**

Fig. 3. **Targeted Data Augmentation.** Undetected tumors are leveraged as informative cases to generate synthetic training samples guided by GPT-4o-produced descriptions, enhancing the model's sensitivity. **Left:** Original false-negative tumors, presented in three orthogonal views (axial, coronal, and sagittal), and their paired radiology reports generated by GPT-4o, describing tumor features like heterogeneous enhancement, cystic appearance, and ill-defined margins. **Right:** Synthetic tumors generated by TextoMorph, conditioned on the original tumor masks and their corresponding GPT-4o-generated reports.

promote semantic consistency and visual fidelity, we sample a similar report $\mathcal{R}_i'$ and generate $\mathcal{T}_i' = g_\theta(\mathcal{R}_i', x_k, m_k)$ under a different CT context (x_k, m_k), minimizing its feature distance to T_i. The contrastive loss is defined as:

$$\mathcal{L}_{\text{contrast}} = \|f(\mathcal{T}_i) - f(\mathcal{T}_i')\|^2 + \max(0, \delta - \|f(\mathcal{T}_i) - f(\mathcal{T}_j)\|)^2. \quad (2)$$

Here, $(\mathcal{T}_i, \mathcal{T}_i')$ are positives (similar text across different CTs) to be minimized, while $(\mathcal{T}_i, \mathcal{T}_j)$ are negatives (distinct texts under the same CT) enforced apart by margin δ, where $f(\cdot)$ is a frozen feature extractor and δ is a predefined margin.

C. Tumor Synthesis

Randomized Tumor Synthesis. We assembled a large repository of healthy CT volumes to ground text- and mask-conditioned tumor synthesis. It contains liver, pancreas, and kidney volumes sampled across scanners and sites to ensure protocol diversity. Tumor masks are seeded by randomized ellipsoids and iteratively adjusted with radiologist feedback for realism [4]. Textual report $\mathcal{R}_{\text{rand}}$ is sampled from a curated pool of 34,035 de-identified radiology reports, providing broad coverage of texture and margin expressions. Combining the healthy context, randomized masks, and sampled reports yields realistic, heterogeneous syntheses that enhance downstream detection, segmentation, and classification. Concretely, given a healthy CT, x_{rand} from the repository, a randomized mask m_{rand}, and a sampled report $\mathcal{R}_{\text{rand}}$, we synthesize:

$$\mathcal{T}^{\text{rand}} = g_\theta(\mathcal{R}_{\text{rand}}, x_{\text{rand}}, m_{\text{rand}}). \quad (3)$$

Targeted Tumor Synthesis. Targeted data augmentation addresses false negative (FN) tumors encountered in detection and segmentation tasks, i.e., real tumors missed by the model, to enhance the model's ability to recognize complex and rare tumors. Given FN cases $\{(\mathcal{T}^{\text{FN}}, m_{\text{FN}})\}_{\text{FN}=1}^{n}$ collected from prior detectors, we use GPT-4o [44] with few-shot prompting to derive concise descriptors $\mathcal{R}_{\text{FN}}$ (texture, margins, etc.) for each $(\mathcal{T}^{\text{FN}}, m_{\text{FN}})$. The text-driven 3D diffusion *model* (§III-B) conditions on $\mathcal{R}_{\text{FN}}$, the tumor mask m_{FN}, and a healthy CT x_{rand}, the text and mask specify appearance and location, while x_{rand} provides normal context. Iterative latent denoising then synthesizes realistic hard examples resembling FN cases, expanding coverage of rare/complex phenotypes and improving downstream detection/segmentation performance. Formally, for each $(x_{\text{FN}}, m_{\text{FN}})$ we generate:

$$\mathcal{T}^{\text{FN}'} = g_\theta(\mathcal{R}_{\text{FN}}, x_{\text{rand}}, m_{\text{FN}}). \quad (4)$$

Class-specific Tumor Synthesis. We leveraged class information to drive category-specific augmentation. For example, pancreatic classes (e.g., PDAC) are distilled into large pools of concise, class-discriminative textual descriptors and paired with large sets of randomized masks. A frozen diffusion generator, jointly conditioned on the mask and text, synthesizes class-consistent tumors at scale. The resulting class-balanced additions sharpen decision boundaries and mitigate distribution shift, improving classification performance. Concretely, for a target class with a class-consistent text $\mathcal{R}_{cls}$, a healthy CT x_{rand}, and randomized mask m_{rand}, we produce:

$$\mathcal{T}^{\text{cls}} = g_\theta(\mathcal{R}_{cls}, x_{\text{rand}}, m_{\text{rand}}). \quad (5)$$

D. Segmentation Model Training

Tumor Detection & Segmentation. We trained a 3D segmentation model on the synthesized tumor data generated under joint mask–text conditioning by our frozen diffusion model. The network architecture and training protocol follow the same design as DiffTumor [9]. The model learns voxel-wise tumor delineation from synthetic and real cases jointly, enabling quantitative assessment of how text-guided synthesis improves lesion boundary precision and small-lesion recall.

Tumor Classification. For tumor type classification, we adopt the same 3D backbone as the segmentation model but modify the output head to predict a scalar probability rather than a

TABLE I

TUMOR REALISM: TEXT-DRIVEN VISUAL TURING TEST (§IV-B). WE REPORT THE READER DISCRIMINATION RATE (%), DEFINED AS THE ACCURACY OF DISTINGUISHING REAL FROM SYNTHETIC TUMORS. RESULTS ARE REPORTED AS SENIOR | JUNIOR READER. LOWER VALUES INDICATE HIGHER REALISM. EACH EVALUATION INCLUDED 60 TUMORS: 20 REAL, 20 GENERATED BY BASELINE METHODS, AND 20 GENERATED BY TEXTOMORPH. BY CONDITIONING ON BOTH TUMOR MASKS AND RADIOLOGY REPORTS, TEXTOMORPH PRODUCES MORE REALISTIC TUMORS THAN MASK-ONLY APPROACHES [9].

		reader discrimination rate (%)		
tumor synthesis	diameter (mm)	liver	pancreas	kidney
SynTumor [4]	all size	71.0 | 26.5	-	-
Pixel2Cancer [6]	all size	68.4 | 60.9	72.4 | 57.1	75.8 | 67.6
DiffTumor [9]	$d < 20$	80.0 | 72.5	70.0 | 62.5	74.5 | 65.0
	$20 \leq d < 50$	74.5 | 70.0	75.0 | 72.5	77.5 | 75.0
	$d \geq 50$	92.5 | 87.5	75.0 | 67.5	80.0 | 72.5
TextoMorph	$d < 20$	67.5 | 60.0	60.0 | 52.5	67.5 | 57.5
	$20 \leq d < 50$	62.5 | 55.0	60.0 | 55.0	60.0 | 50.0
	$d \geq 50$	77.5 | 75.0	55.0 | 47.5	55.0 | 52.5

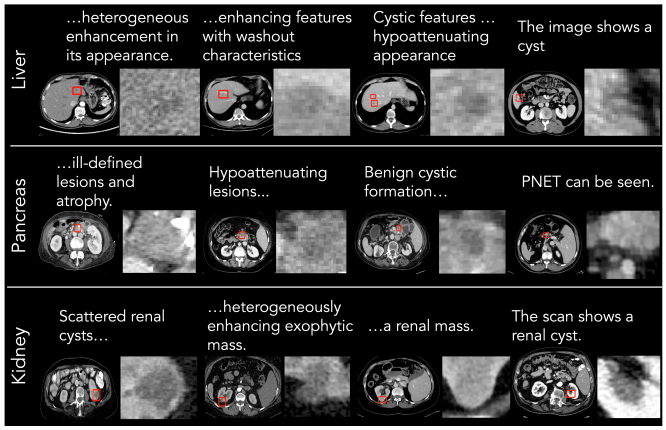


Fig. 4. **Turing Test Visualization (§IV-B).** Examples of synthetic tumors in the liver, pancreas, and kidney, generated using corresponding textual descriptions as input. By jointly conditioning on radiology report text and tumor masks, our method enables precise control over tumor appearance and location, producing tumors that better align with clinical descriptions than mask-only approaches [9].

voxel mask. Specifically, the final convolution and softmax layers are replaced with a global average pooling and a linear classifier that outputs tumor-type probabilities (e.g., cyst vs. PDAC). This configuration preserves spatial feature learning from the segmentation backbone while enabling lesion-level classification. All other training settings (optimizer, learning rate, augmentations) remain identical for consistency.

IV. EXPERIMENT AND RESULT

A. Dataset and Evaluation Metrics

Tumor Synthesis. The training dataset consists of 207 carefully curated CT scans, including 98 liver, 31 pancreas, and 78 kidney cases, each associated with regions of uncertain or lesion status. Ground truth labels indicating tumor presence were initially derived solely from expert-reviewed radiology reports. Since the original dataset did not contain segmentation masks, we employed DiffTumor [9] to semi-automatically generate corresponding tumor masks. From these generated results, we selected a high-confidence subset of true positive cases—66 liver, 15 pancreas, and 60 kidney scans—yielding a total of 141 CT scans paired with reports as input. Additionally, a large-scale corpus comprising 34,035 radiology reports is leveraged in our Text-Driven Contrastive Learning framework to further enhance the synthetic tumor generation process. To evaluate the ***realism*** of synthetic tumors, we measure precision, defined as the reader discrimination rate when distinguishing real from synthetic tumors. Lower reader discrimination rate indicates greater realism in the synthesized tumors. Furthermore, comprehensive radiomic pattern analysis is utilized to quantify tumor ***variability*** by computing the mean variance (MV) of 102 carefully selected radiomic texture features. Higher MV values indicate greater diversity, reflecting enhanced tumor heterogeneity across varied clinical contexts.

Tumor Detection & Segmentation. We used LiTS [81] (131 CTs) for liver, MSD-Pancreas [82] (281 CTs) for pancreas, and KiTS [83] (300 CTs) for kidney to train and test our segmentation models with a 5-fold cross-validation strategy. For healthy data, we selected CT scans from the AbdomenAtlas [32], [35], [36] focusing on the liver, kidney, and pancreas. Tumors were synthesized by randomly deforming spherical shapes to generate masks representing tiny, small, or medium-sized tumors, with randomly selected masks and descriptive text inputs to ensure the realism and variability. For evaluation, we measured detection Sensitivity across small (d<20 mm), medium (20≤d<50 mm), and large (d≥50 mm) tumor sizes, and assessed segmentation quality using the Dice Similarity Coefficient and Normalized Surface Distance.

Tumor Classification. A proprietary dataset comprising 5,119 CT volumes has been utilized in this study, including both normal cases and cases categorized into pancreatic ductal adenocarcinoma (PDAC), cysts, and pancreatic neuroendocrine tumors (PNET) [29] (seeexamples in Figure 7). The data are split into 3,159 training scans and 1,960 test scans. From the training set, we select 20 cases per tumor type and 60 healthy cases to fine-tune our method, and use an additional 120 real tumors to train the classification model. Our study is the *first* to generate class-specific synthetic tumors (PDAC, PNET, cysts). We evaluate performance at the patient level using Sensitivity and Positive Predictive Value (PPV).

B. Tumor Realism: Visual Turing Test

In this Visual Turing Test, two independent readers (Senior and Junior) evaluated 540 CT scans in a randomized, blinded setting to determine the ability to distinguish real tumors from synthetic ones. The scans were categorized by organ type—liver, pancreas, and kidney—and divided into three tumor size ranges: small ($d < 20$ mm), medium ($20 \leq d < 50$ mm), and large ($d \geq 50$ mm). For each category, 60 scans were assessed (20 real, 20 DiffTumor [9]-generated, 20 TextoMorph-generated). Each synthetic case reused the mask of a corresponding real tumor; TextoMorph additionally conditioned on the paired radiology report to better preserve anatomical placement and textural descriptors. Additional comparisons were conducted with methods such as SynTumor [4] and Pixel2Cancer [6].

TABLE II

TUMOR VARIABILITY: RADIOMICS PATTERN ANALYSIS (§IV-C). MEAN VARIANCE (MV) OF TEXTURE-RELATED RADIOMICS FEATURES FOR SYNTHETIC TUMORS GENERATED BY BASELINE METHODS AND TEXTOMORPH ACROSS LIVER, PANCREAS, AND KIDNEY. HIGHER MV INDICATES GREATER TUMOR DIVERSITY.

	mean variance (MV)		
methods	liver	pancreas	kidney
cGANs [84]	1.00	0.89	0.92
SynTumor [4]	1.03	0.92	0.89
Pixel2Cancer [6]	0.99	0.91	1.00
DiffTumor [9]	1.09	**0.95**	0.93
LeFusion [85]	1.10	0.92	1.01
TextoMorph	**1.14**	0.94	**1.03**

The readers provided case-level labels (real versus synthetic). We report the discrimination rate in percent for the Senior and Junior readers respectively; lower values indicate higher realism. As summarized in Table I, TextoMorph shows clear reductions versus DiffTumor across organs and sizes. For large liver lesions, DiffTumor yielded 92.5% and 87.5% discrimination rate, while TextoMorph reduced this to 77.5% and 75.0%. For large pancreas tumors, DiffTumor scored 75.0% and 67.5%, whereas TextoMorph dropped to 55.0% and 47.5%. For large kidney tumors, TextoMorph similarly lowered the rates to 55.0% and 52.5%. Against all-size baselines, TextoMorph is also lower than SynTumor in liver and lower than Pixel2Cancer in both pancreas and kidney. Overall, the Junior reader consistently exhibited lower discrimination rates, yet the relative improvement of TextoMorph over baselines remains robust across both readers.

C. Tumor Variability: Radiomics Pattern Analysis

To assess the diversity of generated tumor appearances, we build upon previous studies [86]–[88] and conduct a radiomics pattern analysis to evaluate synthetic tumors produced by various methods [4], [6], [9], [85]. Specifically, we analyze variance in texture radiomics features [89] to measure how well each model captures tumor heterogeneity. We extract radiomics features (intensity and texture) from 720 synthetic tumors (120 per method; 40 per organ). For a fair comparison, all methods use the same tumor masks and healthy CT scans as spatial and background constraints.

We quantify tumor diversity by the mean variance (MV); higher MV indicates greater heterogeneity. As shown in Table II, we compare MV across liver, pancreas, and kidney for TextoMorph and strong baselines. Relative to stronger baselines, TextoMorph shows higher MV than DiffTumor on liver (1.14 vs. 1.09) and kidney (1.03 vs. 0.95), and exceeds LeFusion on liver (1.14 vs. 1.10) and kidney (1.03 vs. 1.01); pancreas is comparable to DiffTumor (0.94 vs. 0.95) while surpassing LeFusion (0.94 vs. 0.92).

D. Tumor Detection & Segmentation

Figure 5 and Table III compare TextoMorph with state-of-the-art methods [4], [6], [9], [85] across multiple organs and metrics. We then conduct targeted ablation studies to isolate the impact of each core component, and further evaluate the model's generalizability across diverse patient demographics.

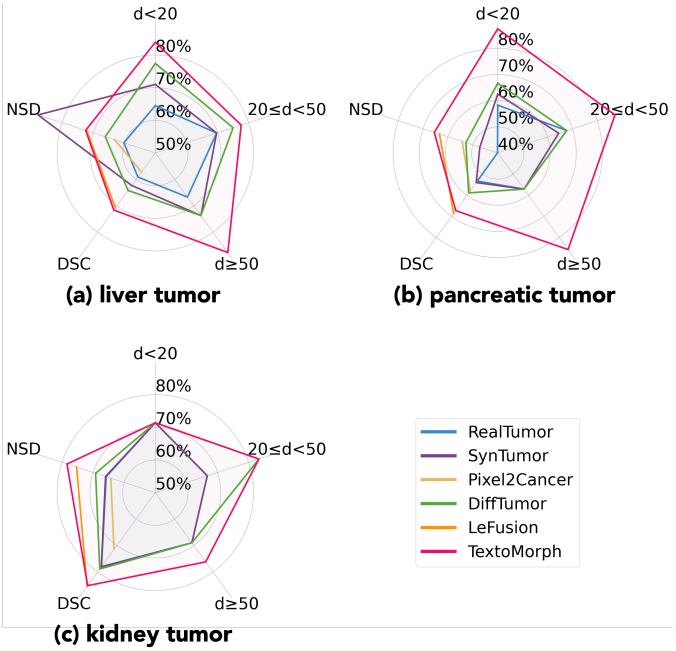


Fig. 5. **Tumor Detection & Segmentation (§IV-D).** We compare the performance of TextoMorph with existing generative methods, measured by sensitivity for detecting small (d<20 mm), medium (20≤d<50 mm), and large (d≥50 mm) tumors, Dice Similarity Coefficient (DSC), and Normalized Surface Distance (NSD).

Ablation on Tumor Report Preprocessing (§III-A). To evaluate the effect of Text Extraction and Augmentation, we compare TextoMorph with and without text augmentation. In the version without text augmentation, only discrete and complex radiology reports are used for tumor generation. Experimental results indicate that the version without text augmentation fails to significantly improve the AI's ability to segment and detect tumors. Specifically, as shown in Table III, for large tumors ($d \geq 50$ mm), the detection rate remains at 75.4%, highlighting its limited capability in handling challenging cases.

Ablation on Text-Driven Contrastive Learning (§III-B). To study the impact of contrastive learning, we compare TextoMorph with and without the contrastive loss function. In this approach, the model maximizes similarity between tumors generated from similar radiology reports while increasing dissimilarity between those generated from distinct reports. This encourages the model to better capture subtle variations in tumor morphology, enhancing its ability to differentiate tumor types and improve segmentation accuracy for complex tumors, such as those with irregular borders or mixed densities. As demonstrated in Table III, increases the detection rate for large liver tumors by 3.5% and improves the NSD by 1.6%.

Ablation on Targeted Data Synthesis (§III-C). To address the limitations of prior tumor detection methods, we introduce Targeted Data Augmentation by leveraging False Negative (FN) tumors as shown in Figure 3. These challenging examples are magnified and paired with descriptive text generated by GPT-4o based on tumor-specific terminology. This structured

TABLE III

ABLATION STUDY ON THE THREE PROPOSED COMPONENTS (§IV-D): COMPARISON OF SENSITIVITY, DICE SIMILARITY COEFFICIENT (DSC), AND NORMALIZED SURFACE DISTANCE (NSD) FOR LIVER, PANCREAS, AND KIDNEY TUMORS USING SYNTHETIC DATA FOR TRAINING WITH U-NET.

Method	Proposed Component			Sensitivity (%) wrt. Tumor Size (d, mm)			DSC (%)	NSD (%)
	§III-A	§III-B	§III-C	$d < 20$	$20 \leq d < 50$	$d \geq 50$		
Liver Tumors								
RealTumor	-	-	-	64.5 (20/31)	69.7 (53/76)	66.7 (38/57)	59.1±30.4	60.1±30.0
SynTumor [4]	-	-	-	71.0 (22/31)	69.7 (53/76)	73.7 (42/57)	62.3±12.7	87.7±21.4
Pixel2Cancer [6]	-	-	-	-	-	-	57.2±21.3	63.1±15.6
DiffTumor [9]	-	-	-	77.4 (24/31)	75.0 (57/76)	73.7 (42/57)	64.2±33.3	66.1±32.8
LeFusion [85]	-	-	-	-	-	-	70.8±9.1	72.0±13.1
TextoMorph (ours)	✗	✗	✗	74.2 (23/31)	72.4 (55/76)	75.4 (43/57)	65.5±25.0	61.3±28.6
	✓	✗	✗	77.4 (24/31)	75.0 (57/76)	77.2 (44/57)	68.4±30.4	69.2±31.0
	✓	✓	✗	80.6 (25/31)	77.6 (59/76)	80.7 (46/57)	69.7±27.2	70.8±26.0
	✓	✓	✓	83.9 (26/31)	77.6 (59/76)	87.7 (50/57)	71.6±27.2	72.4±30.3
Pancreatic Tumors								
RealTumor	-	-	-	58.3 (14/24)	67.7 (21/31)	57.1 (4/7)	53.3±28.7	40.1±28.8
SynTumor [4]	-	-	-	62.5 (15/24)	64.5 (20/31)	57.1 (4/7)	54.0±31.4	47.2±23.0
Pixel2Cancer [6]	-	-	-	-	-	-	57.9±13.7	54.3±19.2
DiffTumor [9]	-	-	-	66.7 (16/24)	67.7 (21/31)	57.1 (4/7)	58.9±42.8	52.8±26.2
LeFusion [85]	-	-	-	-	-	-	68.7±13.5	63.4±21.0
TextoMorph (ours)	✗	✗	✗	66.7 (16/24)	64.5 (20/31)	57.1 (4/7)	55.8±32.6	51.1±35.6
	✓	✗	✗	70.8 (17/24)	61.3 (19/31)	57.1 (4/7)	59.7±36.1	60.6±38.3
	✓	✓	✗	66.7 (16/24)	67.7 (21/31)	57.1 (4/7)	60.2±27.3	71.0±31.5
	✓	✓	✓	87.5 (21/24)	87.1 (27/31)	85.7 (6/7)	67.3±24.8	65.5±27.1
Kidney Tumors								
RealTumor	-	-	-	71.4 (5/7)	66.7 (4/6)	69.0 (29/42)	78.0±14.9	65.8±17.7
SynTumor [4]	-	-	-	71.4 (5/7)	66.7 (4/6)	69.0 (29/42)	78.1±23.0	66.0±21.2
Pixel2Cancer [6]	-	-	-	-	-	-	71.5±21.4	64.3±16.9
DiffTumor [9]	-	-	-	71.4 (5/7)	83.3 (5/6)	69.0 (29/42)	78.9±19.7	69.2±18.5
LeFusion [85]	-	-	-	-	-	-	85.5±15.1	75.3±19.9
TextoMorph (ours)	✗	✗	✗	57.1 (4/7)	83.3 (5/6)	69.0 (29/42)	79.2±22.3	71.4±21.4
	✓	✗	✗	71.4 (5/7)	83.3 (5/6)	76.2 (32/42)	80.6±21.8	76.8±19.3
	✓	✓	✗	71.4 (5/7)	83.3 (5/6)	73.8 (31/42)	79.7±20.2	75.2±21.5
	✓	✓	✓	71.4 (5/7)	83.3 (5/6)	76.2 (32/42)	85.2±9.7	78.4±13.9

input, including tumor masks and healthy CT scans, serves as control conditions for tumor synthesis using a diffusion model. Relative to the RealTumor baseline, DSC increases by 7.2% and sensitivity by 7.2% for kidney large tumor.

E. Generalizable to Different Demographics

To evaluate the enhancement provided by TextoMorph across demographics, we used a proprietary dataset [27], [29], [89], [90] containing malignant pancreatic tumors (PDACs) from patients of varying ages and genders (Figure 6).

Since TextoMorph is designed with an early-screening goal, our synthesis prioritizes *small* ($d < 20$ mm) and *malignant* lesions, enabling better recognition of subtle, low-contrast tumors that conventional models often miss. Overall, TextoMorph improves tumor-wise Sensitivity from 61.9% to 70.1% and DSC from 28.1% to 45.5% across all demographics. By age, gains are most pronounced in the younger brackets e.g., (30–40] and (40–50] with Sensitivity rising by about 15% and DSC by 27.4%, because our synthesis explicitly targets small malignant lesions that are more prevalent and harder to detect in these cohorts. For sex subgroups, Sensitivity increases from 53.8% to 61.5% for males, and 39.2% to 59.5% for females, confirming that TextoMorph consistently improves detection of subtle malignant lesions across both groups without subgroup-specific tuning.

TABLE IV

TUMOR CLASSIFICATION PERFORMANCE (§IV-F). TUMOR-LEVEL SENSITIVITY AND POSITIVE PREDICTIVE VALUE (PPV) FOR MALIGNANT TUMORS (PDAC, PNET) AND BENIGN CYSTS ON A PROPRIETARY TEST DATASET [29]. WE COMPARE REALTUMOR (TRAINED ON REAL DATA ONLY) WITH TEXTOMORPH, WHICH AUGMENTS REAL DATA USING CLASS-SPECIFIC SYNTHETIC TUMORS.

method	malignant tumor		benign cyst	
	sensitivity	PPV	sensitivity	PPV
RealTumor	61.9 (304/491)	46.1 (390/846)	50.7 (245/483)	27.4 (261/953)
TextoMorph	70.1 (344/491)	55.2 (359/650)	57.8 (279/483)	43.1 (286/663)

F. Benign & Malignant Tumor Classification

TextoMorph generates class-specific tumors to improve tumor-level classification and model robustness. For pancreatic tumors, we consider three classes: PDAC, PNET, and cysts. For each class, we select 20 representative CT scans and derive concise, class-consistent textual descriptions (e.g., *'a cystic lesion in the pancreas is present'*). These descriptions, together with the corresponding CT scans and tumor masks, are used to fine-tune the text-driven 3D diffusion model for reliable class-conditioned tumor synthesis.

Using the fine-tuned model, we generate 40 additional synthetic tumors per class. The synthetic tumors are combined with real tumor cases and healthy scans to form a balanced training set for classification. The RealTumor baseline is

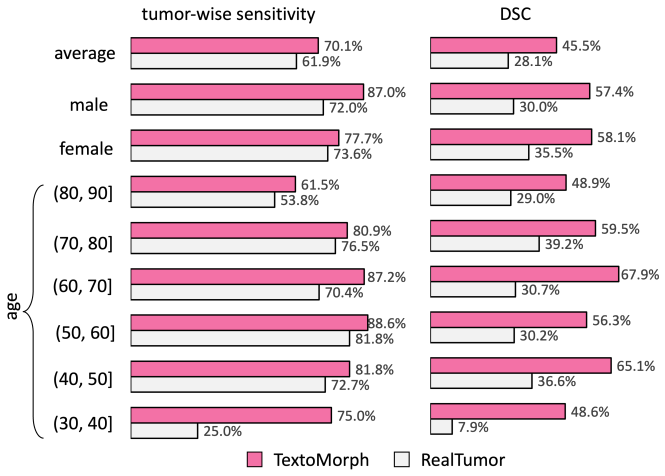


Fig. 6. **Generalizable Across Patient Demographics (§IV-E).** TextoMorph demonstrates consistent performance improvements in detecting malignant tumors in the pancreas (e.g., PDAC) in both tumor-wise Sensitivity (%) and segmentation DSC (%) across various patient demographic groups.

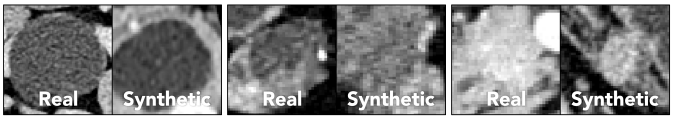


Fig. 7. **Tumor Sub-Type Synthesis.** Visual comparison of real and synthetic pancreatic tumors across three clinically important subtypes: cyst, PDAC, and PNET. The left panel shows real images with characteristic appearances (fluid-filled cysts, hypoattenuating PDAC, and hypervascular PNET), while the right panel shows the corresponding synthetic tumors, demonstrating accurate preservation of subtype-specific texture, shape, and contrast.

trained on real data only, whereas TextoMorph is trained on the same real data augmented with class-specific synthetic tumors.

As shown in Table IV, augmenting with TextoMorph yields consistent gains. For malignant tumors, RealTumor achieves 46.1% PPV, whereas TextoMorph improves these to 55.2% PPV (+9.1%). For benign cysts, RealTumor attains 50.7% sensitivity, while TextoMorph increases them to 57.8% sensitivity (+7.1%). These results indicate that class-aware synthetic tumors generated by TextoMorph strengthen both detection and discrimination of pancreatic tumor types. Additional tumor- and patient-level comparisons are provided in Figure 7.

V. Conclusion

TextoMorph improves AI for cancer imaging by generating realistic, diverse tumors with fine-grained control over key characteristics in CT scans, such as texture, boundaries, size, and pathology. By exploiting descriptive text from radiology reports, TextoMorph addresses the limitations of existing synthesis methods, enabling targeted data augmentation to create tumors that AI models often miss due to the scarcity of training CT scans with real tumors, leading to significant improvements in tumor detection, segmentation, and classification. Furthermore, this text-driven synthesis reduces reliance on scarce annotated medical datasets, offering a scalable and efficient solution to augment medical imaging data and better address critical clinical needs.

Acknowledgment

We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank Wenxuan Li and Jieneng Chen for their helpful suggestions throughout the project. Paper content is covered by patents pending.

References

[1] J. Niemeijer, J. Ehrhardt, H. Uzunova, and H. Handels, "Tsynd: Targeted synthetic data generation for enhanced medical image classification: Leveraging epistemic uncertainty to improve model performance," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2024, pp. 69–78.

[2] B. D. Basaran, W. Zhang, M. Qiao, B. Kainz, P. M. Matthews, and W. Bai, "Lesionmix: A lesion-level data augmentation method for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 73–83.

[3] Q. Chen, Y. Lai, X. Chen, Q. Hu, A. Yuille, and Z. Zhou, "Analyzing tumors by synthesis," *Generative Machine Learning Models in Medical Image Computing*, pp. 85–110, 2024.

[4] Q. Hu, Y. Chen, J. Xiao, S. Sun, J. Chen, A. L. Yuille, and Z. Zhou, "Label-free liver tumor segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7422–7432. [Online]. Available: https://github.com/MrGiovanni/SyntheticTumors

[5] Q. Hu, A. Yuille, and Z. Zhou, "Synthetic data as validation," *arXiv preprint arXiv:2310.16052*, 2023. [Online]. Available: https://github.com/MrGiovanni/SyntheticValidation

[6] Y. Lai, X. Chen, A. Wang, A. Yuille, and Z. Zhou, "From pixel to cancer: Cellular automata in computed tomography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2024, pp. 36–46.

[7] Q. Chen, X. Zhou, C. Liu, H. Chen, W. Li, Z. Jiang, Z. Huang, Y. Zhao, D. Yu, J. He, Y. Zheng, L. Shao, A. Yuille, and Z. Zhou, "Scaling tumor segmentation: Best lessons from real and synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24 001–24 013. [Online]. Available: https://github.com/BodyMaps/AbdomenAtlas2.0

[8] B. Gonçalves, M. Silva, L. Vieira, and P. Vieira, "Abdominal mri unconditional synthesis with medical assessment," *BioMedInformatics*, vol. 4, no. 2, pp. 1506–1518, 2024.

[9] Q. Chen, X. Chen, H. Song, Z. Xiong, A. Yuille, C. Wei, and Z. Zhou, "Towards generalizable tumor synthesis," in *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2024, pp. 11 147–11 158. [Online]. Available: https://github.com/MrGiovanni/DiffTumor

[10] P. R. Bassi, X. Zhou, W. Li, S. Płotka, J. Chen, Q. Chen, Z. Zhu, J. Prządo, I. E. Hamamci, S. Er, X. Chen, M. C. Yavuz, Y.-C. Chou, T. Lin, K. Wang, Y. Tang, J. B. Cwikla, S. Decherchi, A. Cavalli, Y. Yang, A. L. Yuille, and Z. Zhou, "Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks," *arXiv preprint arXiv:2510.14803*, 2025. [Online]. Available: https://github.com/MrGiovanni/R-Super

[11] P. R. Bassi, W. Li, J. Chen, Z. Zhu, T. Lin, S. Decherchi, A. Cavalli, K. Wang, Y. Yang, A. L. Yuille, and Z. Zhou, "Learning segmentation from radiology reports," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 305–315. [Online]. Available: https://github.com/MrGiovanni/R-Super

[12] P. R. Bassi, M. C. Yavuz, I. E. Hamamci, S. Er, X. Chen, W. Li, B. Menze, S. Decherchi, A. Cavalli, K. Wang, Y. Yang, A. Yuille, and Z. Zhou, "Radgpt: Constructing 3d image-text tumor datasets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 23 720–23 730. [Online]. Available: https://github.com/MrGiovanni/RadGPT

[13] Z. Zhou, M. B. Gotway, and J. Liang, "Interpreting medical images," in *Intelligent Systems in Medicine and Health*. Springer, 2022, pp. 343–371.

[14] Z. Zhou, "Towards annotation-efficient deep learning for computer-aided diagnosis," Ph.D. dissertation, Arizona State University, 2021. [Online]. Available: https://github.com/MrGiovanni/Dissertation

[15] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[16] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[17] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[18] P. Guo, C. Zhao, D. Yang, Y. He, V. Nath, Z. Xu, P. R. Bassi, Z. Zhou, B. D. Simon, S. A. Harmon, A. B. Syed, H. Roth, and D. Xu, "Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model," *arXiv preprint arXiv:2505.04522*, 2025.

[19] J. Mao, Y. Wang, Y. Tang, D. Xu, K. Wang, Y. Yang, Z. Zhou, and Y. Zhou, "Medsegfactory: Text-guided generation of medical image-mask pairs," *arXiv preprint arXiv:2504.06897*, 2025. [Online]. Available: https://github.com/jwmao1/MedSegFactory

[20] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

[21] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in neural information processing systems*, vol. 35, pp. 8633–8646, 2022.

[22] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[23] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 563–22 575.

[24] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[25] N. Heller, F. Isensee, D. Trofimova *et al.*, "The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct," 2023.

[26] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2015, pp. 556–564.

[27] Y.-C. Chou, Z. Zhou, and A. Yuille, "Embracing massive medical data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 24–35. [Online]. Available: https://github.com/MrGiovanni/OnlineLearning

[28] P. R. Bassi, W. Li, Y. Tang, F. Isensee, Z. Wang, J. Chen, Y.-C. Chou, Y. Kirchhoff, M. Rokuss, Z. Huang, J. Ye, J. He, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. H. Maier-Hein, P. Jaeger, Y. Ye, Y. Xie, J. Zhang, Z. Chen, Y. Xia, Z. Xing, L. Zhu, Y. Sadegheih, A. Bozorgpour, P. Kumari, R. Azad, D. Merhof, P. Shi, T. Ma, Y. Du, F. Bai, T. Huang, B. Zhao, H. Wang, X. Li, H. Gu, H. Dong, J. Yang, M. A. Mazurowski, S. Gupta, L. Wu, J. Zhuang, H. Chen, H. Roth, D. Xu, M. B. Blaschko, S. Decherchi, A. Cavalli, A. L. Yuille, and Z. Zhou, "Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?" *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, pp. 15 184–15 201, 2024. [Online]. Available: https://github.com/MrGiovanni/Touchstone

[29] Y. Xia, Q. Yu, L. Chu, S. Kawamoto, S. Park, F. Liu, J. Chen, Z. Zhu, B. Li, Z. Zhou, A. L. Yuille, E. K. Fishman, and R. H. Hruban, "The felix project: Deep networks to detect pancreatic neoplasms," *medRxiv*, 2022.

[30] A. Lubonja, P. R. Bassi, W. Li, H. Qiao, R. Burns, A. L. Yuille, and Z. Zhou, "Auditing significance, metric choice, and demographic fairness in medical ai challenges," *arXiv preprint arXiv:2512.19091*, 2025.

[31] W. Li, X. Zhou, Q. Chen, T. Lin, P. R. Bassi, S. Plotka, J. B. Cwikla, X. Chen, C. Ye, Z. Zhu, Y.-C. Chou, K. Wang, Y. Tang, A. L. Yuille, and Z. Zhou, "Pants: The pancreatic tumor segmentation dataset," *arXiv preprint arXiv:2507.01291*, 2025. [Online]. Available: https://github.com/MrGiovanni/PanTS

[32] W. Li, C. Qu, X. Chen, P. R. Bassi, Y. Shi, Y. Lai, Q. Yu, H. Xue, Y. Chen, X. Lin, Y. Tang, Y. Cao, H. Han, Z. Zhang, J. Liu, T. Zhang, Y. Ma, J. Wang, G. Zhang, A. Yuille, and Z. Zhou, "Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking," *Medical Image Analysis*, p. 103285, 2024. [Online]. Available: https://github.com/MrGiovanni/AbdomenAtlas

[33] W. Li, P. R. Bassi, T. Lin, Y.-C. Chou, X. Zhou, Y. Tang, F. Isensee, K. Wang, Q. Chen, X. Xu, J. Ye, Z. Zhu, S. Decherchi, A. Cavalli, A. L. Yuille, and Z. Zhou, "Scalemai: Accelerating the development of trusted datasets and ai models," *arXiv preprint arXiv:2501.03410*, 2025. [Online]. Available: https://github.com/MrGiovanni/ScaleMAI

[34] J. Li, Z. Zhou, J. Yang, A. Pepe, C. Gsaxner, G. Luijten, C. Qu, T. Zhang, X. Chen, W. Li, Y. Jin, and J. Egger, "Medshapenet–a large-scale dataset of 3d medical shapes for computer vision," *Biomedical Engineering/Biomedizinische Technik*, no. 0, 2024. [Online]. Available: https://medshapenet.ikim.nrw

[35] C. Qu, T. Zhang, H. Qiao, J. Liu, Y. Tang, A. Yuille, and Z. Zhou, "Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks," in *Conference on Neural Information Processing Systems (NeurIPS)*, vol. 21, 2023. [Online]. Available: https://github.com/MrGiovanni/AbdomenAtlas

[36] W. Li, A. Yuille, and Z. Zhou, "How well do supervised models transfer to 3d image segmentation?" in *International Conference on Learning Representations*, 2024. [Online]. Available: https://github.com/MrGiovanni/SuPreM

[37] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[38] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[39] Q. Hu, J. Xiao, Y. Chen, S. Sun, J.-N. Chen, A. Yuille, and Z. Zhou, "Synthetic tumors make ai segment tumors better," *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. [Online]. Available: https://github.com/MrGiovanni/SyntheticTumors

[40] S. Du, X. Wang, Y. Lu, Y. Zhou, S. Zhang, A. Yuille, K. Li, and Z. Zhou, "Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5. [Online]. Available: https://github.com/zoedsy/BoostingLesionSegviaDiffusionModels

[41] B. Li, Y.-C. Chou, S. Sun, H. Qiao, A. Yuille, and Z. Zhou, "Early detection and localization of pancreatic cancer by label-free tumor synthesis," *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023. [Online]. Available: https://github.com/MrGiovanni/SyntheticTumors

[42] Y. Xu, L. Sun, W. Peng, S. Jia, K. Morrison, A. Perer, A. Zandifar, S. Visweswaran, M. Eslami, and K. Batmanghelich, "Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images," *IEEE Transactions on Medical Imaging*, 2024.

[43] Y. Yang, Z.-Y. Wang, Q. Liu, S. Sun, K. Wang, R. Chellappa, Z. Zhou, A. Yuille, L. Zhu, Y.-D. Zhang, and J. Chen, "Medical world model: Generative simulation of tumor evolution for treatment planning," *arXiv preprint arXiv:2506.02327*, 2025. [Online]. Available: https://github.com/scott-yjyang/MeWM

[44] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[45] P. R. Bassi, Q. Wu, W. Li, S. Decherchi, A. Cavalli, A. Yuille, and Z. Zhou, "Label critic: Design data before models," in *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2025, pp. 1–5. [Online]. Available: https://github.com/PedroRASB/LabelCritic

[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[47] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," *IEEE Winter Conference on Applications of Computer Vision*, 2022. [Online]. Available: https://github.com/lambert-x/medical_mae

[48] Y. Shin, H. A. Qadir, and I. Balasingham, "Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance," *IEEE Access*, vol. 6, pp. 56 007–56 017, 2018.

[49] B. Billot *et al.*, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining," *Medical Image Analy.*, vol. 86, p. 102789, 2023.

[50] C. Han *et al.*, "Synthesizing diverse lung nodules wherever massively: 3d

multi-conditional gan-based ct image augmentation for object detection," in *3DV*. IEEE, 2019, pp. 729–737.
[51] F. Lyu *et al.*, "Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation," *IEEE Trans. Medical Imag.*, vol. 42, no. 3, pp. 797–809, 2022.
[52] Q. Yao, L. Xiao, P. Liu, and S. K. Zhou, "Label-free segmentation of covid-19 lesions in lung ct," *IEEE Trans. Medical Imag.*, vol. 40, no. 10, pp. 2808–2819, 2021.
[53] S. Du *et al.*, "Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts," *arXiv preprint arXiv:2310.02906*, 2023.
[54] J. Wei, Y. Li, M. Qiu, H. Chen, X. Fan, and W. Lei, "Sam-fnet: Sam-guided fusion network for laryngo-pharyngeal tumor detection," *arXiv preprint arXiv:2408.05426*, 2024.
[55] J. Wei, Y. Li, X. Fan, W. Ma, M. Qiu, H. Chen, and W. Lei, "Sam-swin: Sam-driven dual-swin transformers with adaptive lesion enhancement for laryngo-pharyngeal tumor detection," *arXiv preprint arXiv:2410.21813*, 2024.
[56] Q. Jin, H. Cui, C. Sun, Z. Meng, and R. Su, "Free-form tumor synthesis in computed tomography images via richer generative adversarial network," *Knowledge-Based Systems*, vol. 218, p. 106753, 2021.
[57] H. Wang, Y. Zhou, J. Zhang, J. Lei, D. Sun, F. Xu, and X. Xu, "Anomaly segmentation in retinal images with poisson-blending data augmentation," *Medical Image Analysis*, vol. 81, p. 102534, 2022.
[58] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *CVPR*, 2022, pp. 650–656.
[59] L. Yang, X. Xu, B. Kang, Y. Shi, and H. Zhao, "Freemask: Synthetic images with dense annotations make stronger segmentation models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[60] H. Li, Y. Fan, and J. Wu, "Tumor synthesis with adversarial networks for augmenting data in medical imaging," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2380–2390, 2020.
[61] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Lung nodule classification with multi-scale convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 588–595.
[62] J. Doe, J. Smith, and M. Lee, "Mac-dm: Mask-controlled diffusion models for synthetic distal tibial radiographs," *Medical Image Analysis*, vol. 67, p. 101812, 2021.
[63] L. Wu, J. Zhuang, Y. Zhou, S. He, J. Ma, L. Luo, X. Wang, X. Ni, X. Zhong, M. Wu *et al.*, "Large-scale generative tumor synthesis in computed tomography images for improving tumor recognition," *Nature Communications*, vol. 16, no. 1, p. 11053, 2025.
[64] Z. Yang, Z. Chen, Y. Sun, A. Strittmatter, A. Raj, A. Allababidi, J. Rink, and F. G. Zoellner, "seg2med: a bridge from artificial anatomy to multimodal medical images," *Physics in Medicine and Biology*, 2025.
[65] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
[66] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.
[67] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 3876–3887.
[68] L. Wu, J. Zhuang, and H. Chen, "VoCo: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
[69] J. Yang, C. Li, P. Zhang, B. Xiao, C. Liu, L. Yuan, and J. Gao, "Unified contrastive learning in image-text-label space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 163–19 173.
[70] S. Yellapragada, A. Graikos, P. Prasanna, T. Kurc, J. Saltz, and D. Samaras, "Pathldm: Text conditioned latent diffusion model for histopathology," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5182–5191.
[71] M. Hamamci, U. Kantarci, and B. Yaman, "Generatect: Text-guided 3d medical image synthesis for multi-abnormality classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1234–1245, 2023.
[72] X. Chen, Y. Li, and Y. Zhu, "Text2image: Synthesizing chest x-rays from radiology reports using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 10–19.
[73] S. Park, H. Lee, and M. Kang, "Generating retinal images from text descriptions for ophthalmology applications," *Medical Image Analysis*, vol. 64, p. 101741, 2020.
[74] P. R. A. S. Bassi, Q. Wu, W. Li, S. Decherchi, A. Cavalli, A. Yuille, and Z. Zhou, "Label critic: Design data before models," 2024. [Online]. Available: https://arxiv.org/abs/2411.02753
[75] M. Giuffrè, F. Romano, and F. Vitale, "Harnessing synthetic data in healthcare: Applications, challenges, and future directions," *Journal of Medical Systems*, vol. 47, no. 2, p. 45, 2023.
[76] Z. Wu, S. W. Remedios, B. E. Dewey, A. Carass, and J. L. Prince, "Anires2d: anisotropic residual-enhanced diffusion for 2d mr super-resolution," in *Medical Imaging 2024: Clinical and Biomedical Imaging*, vol. 12930. SPIE, 2024, pp. 567–574.
[77] T. Lin, Z. Chen, Z. Yan, W. Yu, and F. Zheng, "Stable diffusion segmentation for biomedical images with single-step reverse process," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Cham: Springer Nature Switzerland, 2024, pp. 656–666.
[78] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239
[79] W. Yao, C. Liu, K. Yin, W. K. Cheung, and J. Qin, "Addressing asynchronicity in clinical multimodal fusion via individualized chest x-ray generation," *arXiv preprint arXiv:2410.17918*, 2024.
[80] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.
[81] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.
[82] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, p. 4128, 2022.
[83] N. Heller, S. McSweeney, M. T. Peterson, S. Peterson, J. Rickman, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, J. 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.
[84] A. Biswas, S. P. Maity, R. Banik, P. Bhattacharya, and J. Debbarma, "Gan-driven liver tumor segmentation: enhancing accuracy in biomedical imaging," *SN Computer Science*, vol. 5, no. 5, p. 652, 2024.
[85] H. Zhang, Y. Liu, J. Yang, S. Wan, X. Wang, W. Peng, and P. Fua, "Lefusion: Controllable pathology synthesis via lesion-focused diffusion models," *arXiv preprint arXiv:2403.14066*, 2024.
[86] W. Zhang, Y. Guo, and Q. Jin, "Radiomics and its feature selection: A review," *Symmetry*, vol. 15, no. 10, p. 1834, 2023.
[87] H. Nasief, W. Hall, C. Zheng, S. Tsai, and L. Wang, "Improving treatment response prediction for chemoradiation therapy of pancreatic cancer using a combination of delta-radiomics and the clinical biomarker ca19-9," *Frontiers in Oncology*, vol. 10, p. 203, 2020.
[88] L. Peng, V. Parekh, P. Huang, D. D. Lin, K. Sheikh, B. Baker, T. Kirschbaum, F. Silvestri, J. Son, A. Robinson *et al.*, "Distinguishing true progression from radionecrosis after stereotactic radiation therapy for brain metastases with machine learning and radiomics," *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1236–1243, 2018.
[89] L. C. Chu, S. Park, S. Kawamoto, D. F. Fouladi, S. Shayesteh, E. S. Zinreich, J. S. Graves, K. M. Horton, R. H. Hruban, A. L. Yuille *et al.*, "Utility of ct radiomics features in differentiation of pancreatic ductal adenocarcinoma from normal pancreatic tissue," *American Journal of Roentgenology*, vol. 213, no. 2, pp. 349–357, 2019.
[90] M. Kang, B. Li, Z. Zhu, Y. Lu, E. K. Fishman, A. Yuille, and Z. Zhou, "Label-assemble: Leveraging multiple datasets with partial labels," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2023, pp. 1–5. [Online]. Available: https://github.com/MrGiovanni/LabelAssemble