3D-LLDM: LABEL-GUIDED 3D LATENT DIFFUSION MODEL FOR IMPROVING HIGH-RESOLUTION SYNTHETIC MR IMAGING IN HEPATIC STRUCTURE SEGMENTATION

Anonymous Authors

Anonymous Institution

ABSTRACT

Deep learning and generative models are advancing rapidly, with synthetic data increasingly being integrated into training pipelines for downstream analysis tasks. However, in medical imaging, their adoption remains constrained by the scarcity of reliable annotated datasets. To address this limitation, we propose 3D-LLDM, a label-guided 3D latent diffusion model that generates high-quality synthetic magnetic resonance (MR) volumes with corresponding anatomical segmentation masks. Our approach utilizes hepatobiliary phase MR images enhanced with Gd-EOB-DTPA contrast agent to derive structural masks for liver, portal vein, hepatic vein, and hepatocellular carcinoma, which subsequently guide volumetric synthesis through a ControlNet-based architecture. Trained on 720 real clinical hepatobiliary phase MR scans from Samsung Medical Center, 3D-LLDM achieves Fréchet Inception Distance (FID) of 28.31 and a 70.9% improvement over GANs and 26.7% over state-of-the-art diffusion baselines. When used for data augmentation, our synthetic volumes boost hepatocellular carcinoma segmentation by up to 11.153% Dice score across five CNN architectures.

***Index Terms*—** Biomedical imaging, Generative models, Latent space modeling, Multi-modal imaging, Deep neural networks.

1. INTRODUCTION

Deep learning has achieved remarkable advancements in extracting relevant patterns from data and making precise decisions, particularly in image-based tasks such as classification and segmentation [1–4]. The performance of deep learning models has rapidly improved, frequently surpassing traditional methods across various domains.

Generative models are developed to capture underlying data distributions and produce new, synthetic images. Recent studies have shown that traditional generative models can be applied to the medical domain, including MR and CT imaging [5–7]. However, their ability to synthesize high-quality medical volumes remains insufficient for practical applications.

A major challenge in this field is the limited availability of high-quality training data. Since deep learning models rely on extensive and diverse datasets for effective training, the shortage of high-quality medical images restricts the accuracy and realism of the generated synthetic volumes [8–10].

To overcome this constraint, we introduce an innovative approach — the Label Guided 3D Latent Diffusion Model (3D-LLDM) — designed to generate high-resolution synthetic volumes along with their corresponding segmentation labels. In the context of abdominal MRI, our model first synthesizes segmentation labels containing the liver and veins, which are then used to train ControlNet [11] as spatial guidance for volume generation. Leveraging these guided synthetic labels, our model produces realistic synthetic datasets, which can be effectively utilized for downstream tasks.

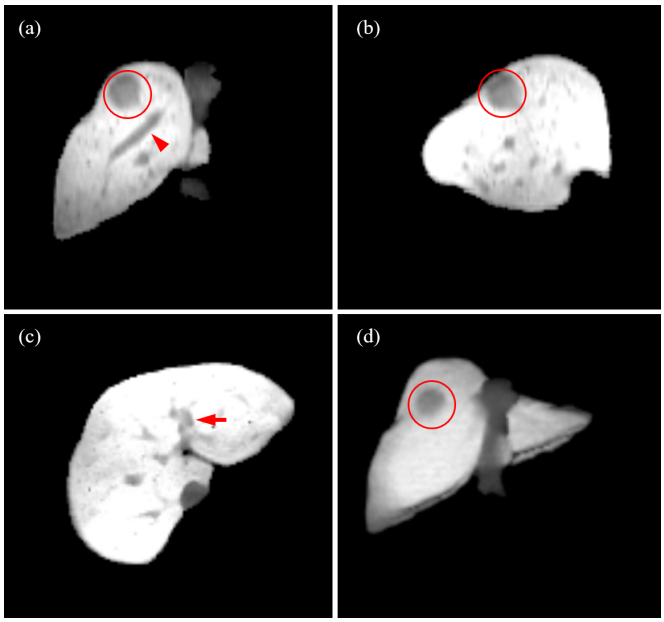

Fig. 1. High-Quality synthetic MR volume generated by 3D-LLDM demonstrating anotomical consistency across multi planes: (a) Coronal image, (b) sagittal image, and (c) axial image reformatted from the high-resolution 3D volumetric data shown in (d). Note the tumor highlighted by a red circle, the right hepatic vein indicated by an arrowhead, and the left portal vein marked by an arrow.

Figure 1 shows coronal, sagittal, and axial views reformatted from 3D volumetric data, highlighting hepatic veins and a tumor. This example illustrates the importance of generating anatomically consistent 3D synthetic volumes.

Our model is trained with the HBP images of 720 Gd-EOB-DTPA-enhanced MR imaging and generates better high-resolution synthetic images compared to conventional models [12]. Additionally, the synthesized volume-label pairs are incorporated into the training datasets for downstream tasks of two different segmentation methods of the hepatic structures including the liver segmentation and multi-class segmentation from the portal vein, hepatic vein and tumor (i.e., hepatocellular carcinoma [HCC]). Comparative results among various CNN-based segmentation models [13, 14] demonstrate that the high-resolution synthetic volumes generated by 3D-LLDM significantly improved the performance of segmentation of the hepatic structures.

Our main contributions are: (1) to present 3D-LLDM, an innovative diffusion based generative model tailored for high-resolution MR volume synthesis, incorporating label guidance to improve generation control and (2) to demonstrate that incorporating 3D-LLDM-generated synthetic data into the training pipeline could improve the performance of CNN-based liver segmentation models, achieving higher accuracy compared to training on real data alone.

2. RELATED WORK

2.1. Diffusion Model

Diffusion-based models have emerged as a powerful generative approach, widely applied in medical imaging tasks such as image synthesis [15–19] due to their capabilities in high-quality image synthesis. Unlike previous approaches, these models excel in generating diagnostically valuable 2D medical images with remarkable detail fidelity. However, extending to 3D remains challenging; current methods like text-guided CT generators create volumes slice-by-slice, resulting in problematic structural inconsistencies between cross sections [20]. Research has increasingly targeted specialized applications, particularly synthetic tumor generation for improving segmentation performance across multiple organs [21–25].

2.2. 3D Latent Diffusion Model

Latent Diffusion Models (LDMs), a variant of diffusion model, operate in a compressed latent space, significantly reducing computational complexity while preserving high-quality outputs [26].

Given a medical volume $x \in \mathbb{R}^{H\times W\times D}$, where H, W, and D represent the dimensions of height, width, and depth, respectively, the encoder $\mathcal{E}$ maps x to a latent representation $z = \mathcal{E}(x) \in \mathbb{R}^{h\times w\times d}$. The decoder $\mathcal{D}$ then reconstructs the volume as $\tilde{x} = \mathcal{D}(z)$ from the latent features. Typically, a Variational Autoencoder (VAE) structure [27] is employed for the encoder-decoder pair $(\mathcal{E}, \mathcal{D})$.

In Latent Diffusion Models (LDMs), the model $\epsilon_\theta(z_t, t)$ is trained to estimate a noise-free reconstruction of the input latent features z_t, where z_t represents a noisy transformation of the original input at time step $t \in [0, T]$. The training objective for the latent diffusion model ϵ_θ is formulated as:

$$\mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t)||_2^2\right]. \tag{1}$$

2.3. ControlNet

ControlNet [11] is a neural network extension that enhances diffusion models by incorporating additional spatial guidance, allowing precise control over the generated output through structured inputs. Combining with a Latent Diffusion Model (LDM), the additional conditioning information is transformed into latent features, denoted as the task-specific condition c_f. The overall learning objective of the diffusion algorithm, incorporating ControlNet ϵ_θ, is formulated as:

$$\mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t,c_f}\left[||\epsilon - \epsilon_\theta(z_t, t, c_f)||_2^2\right]. \tag{2}$$

In this work, we leverage the NVIDIA MONAI framework an open-source AI platform tailored for medical research [28] to effectively adapt ControlNet for medical image processing.

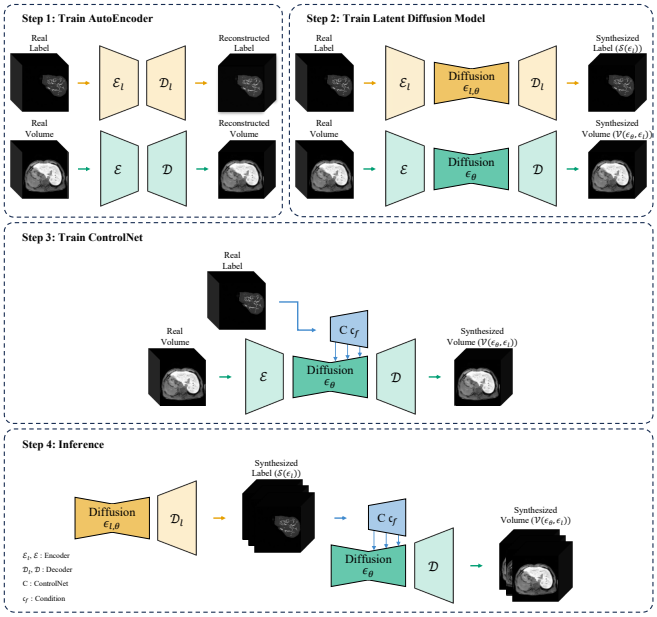

Fig. 2. Overview of the training and inference pipeline of the proposed 3D-LLDM. The process consists of four steps: (1) training the autoencoder for both label and volume reconstruction, (2) training the latent diffusion model for label and volume synthesis, (3) training the ControlNet to condition the diffusion process on labels, and (4) performing inference using the pretrained latent diffusion models and ControlNet.

3. METHODS

3.1. Label-Guided 3D Latent Diffusion Model

In this paper, we propose the Label-Guided 3D Latent Diffusion Model (3D-LLDM) for generating high-resolution synthetic medical volumes paired with corresponding labels. Our approach first trains a synthetic label generation model using a latent diffusion framework. Given the encoder-decoder pair $(\mathcal{E}_l, \mathcal{D}_l)$ for label synthesis, the diffusion model $\epsilon_{l,\theta}$ is optimized using the following learning objective:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\mathcal{E}_l(l),\,\epsilon_l\sim\mathcal{N}(0,1),\,t}\left[\left\|\epsilon_l - \epsilon_{l,\theta}(z_l, t, t)\right\|_2^2\right]. \tag{3}$$

Next, 3D-LLDM utilizes the synthetic labels $L(\epsilon_l) = \mathcal{D}_l(\epsilon_{l,\theta}(\epsilon_l, T))$ as input to ControlNet, which guides the generation of medical volumes. The task-specific condition $c_f(\epsilon_l)$ in the latent space of ControlNet is defined as:

$$c_f(\epsilon_l) = \mathcal{E}\big(\mathcal{D}_l(\epsilon_{l,\theta}(\epsilon_l, T))\big). \tag{4}$$

Then, the learning objective of ControlNet ϵ_θ is formulated as:

$$\mathcal{L}_c = \mathbb{E}_{\mathcal{E}(x),\{\epsilon,\epsilon_l\}\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, c_f(\epsilon_l))||_2^2\right]. \tag{5}$$

In practice, we use the latent-space representations of real labels,z_l, instead of synthetic labels to improve performance, as shown in the following equation:

$$\mathcal{L}_{c,\text{train}} = \mathbb{E}_{\mathcal{E}(x),\{\epsilon_c\}\sim\mathcal{N}(0,1),\,t}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, z_l)\right\|_2^2\right]. \tag{6}$$

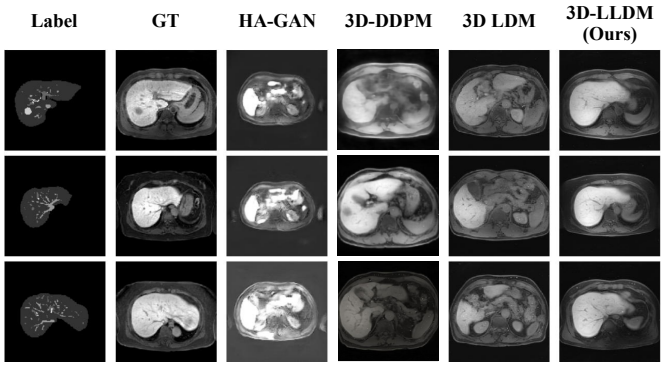

Fig. 3. Qualitative comparison of synthetic MR volumes of various levels generated by state-of-the-art generative models. The columns correspond to: (1) the label-guided input, (2) the ground truth (GT) MRI scan, (3) the HA-GAN-generated synthetic image, (4) the 3D-DDPM-generated synthetic image, (5) the 3D Latent Diffusion Model (3D LDM) with an autoencoder (AE) backbone, and (6) the proposed 3D-LLDM (Ours)-generated synthetic image.

Figure 2 presents the structure and operational process of the proposed 3D-LLDM. The training process consists of two steps. First, our approach trains the autoencoder pairs $(\mathcal{E}_l, \mathcal{D}_l)$ for label synthesis and $(\mathcal{E}, \mathcal{D})$ for volume reconstruction. Subsequently, we train the latent diffusion model $\epsilon_{l,\theta}$ to generate synthetic labels, and ControlNet ϵ_θ to generate synthetic volumes conditioned on the real labels.

The inference process is illustrated in Step 3. Given random noise $\epsilon_l \sim \mathcal{N}(0, 1)$, we first generate synthetic labels as:

$$S(\epsilon_l) = \mathcal{D}_l(\epsilon_{l,\theta}(\epsilon_l, T)). \tag{7}$$

Next, using the generated synthetic labels and additional noise $\epsilon \sim \mathcal{N}(0, 1)$, we synthesize the final medical volume:

$$V(\epsilon_\theta, \epsilon_l) = \mathcal{D}(\epsilon_\theta(z_t, t, c_f(\epsilon_l))). \tag{8}$$

This process produces a paired dataset of synthetic labels $S(\epsilon_l)$ and corresponding synthetic volumes $V(\epsilon_\theta, \epsilon_l)$, which can be effectively leveraged for downstream tasks such as segmentation.

4. EXPERIMENTS

4.1. Dataset and Implementation Details

This study employs anonymized MR imaging using Gd-EOB-DTPA (Bayer HealthCare, Germany) in 720 patients with HCC from Samsung Medical Center (Seoul, Korea). Particularly, we focused on volumes from HBP images, where the hyperintense liver is well discriminated compared to other hypointense abdominal organs. For each patient, we created segmentation labels encompassing the liver, portal vein, hepatic vein, and tumor (i.e., HCC), which serve as the ground truth for segmentation tasks.

To facilitate efficient training and enhance patient privacy protection, we cropped the region of interest (ROI) to a fixed size of $(160, 160, 64)$ before training. After cropping, the data was normalized via range scaling, mapping intensity values to $[0, 1]$.

The dataset was divided into three subsets: 504 patients were designated for model training, 72 for validation purposes, and 144 for final testing evaluation. All computational experiments were performed using a single NVIDIA A100 GPU with 80GB memory, utilizing the MONAI 1.10 framework and Python 3.8 environment. The AdamW optimizer from PyTorch was employed for all gradient-based optimization procedures. The complete training phase for the 3D-LLDM required approximately one week of continuous computation on the A100 hardware.

Method	Ax. FID ↓	Sag. FID ↓	Cor. FID ↓	Avg. FID ↓
HA-GAN	96.32	98.07	97.75	97.38
3D-DDPM	79.53	78.16	76.07	77.92
3D-LDM (w/ VQVAE)	67.17	63.43	62.96	64.52
3D-LDM (w/ VAE)	40.69	37.34	37.83	38.62
3D-LLDM	**29.27**	**27.62**	**28.04**	**28.31**

Table 1. Comparison of FID scores (↓) among different generative models for medical image synthesis. Lower scores indicate higher image quality. The proposed 3D-LLDM model achieves the lowest FID scores in all views (Axial, Sagittal, and Coronal) as well as the overall average, consistently outperforming conventional generative models.

4.2. Evaluation of 3D-LLDM

This study evaluates the quality of generated images using the Fréchet Inception Distance (FID) metric [29]. Feature extraction for FID computation was performed using a 3D ResNet-50 model from Tencent MedicalNet [30], which mapped $160 \times 160 \times 64$ volumetric images to 2048-dimensional feature vectors.

Table 1 compares the proposed 3D-LLDM with HA-GAN, 3D-DDPM, and two variants of 3D latent diffusion models: (1) the default MONAI implementation and (2) a version with a VQ-VAE backbone. 3D-LLDM achieves the lowest FID of **28.31**, representing a **70.9%** relative reduction compared with HA-GAN (FID 97.38) and a **26.7%** relative reduction compared with the strongest diffusion baseline (3D-LDM w/ VAE; FID 38.62).

Figure 3 presents a qualitative comparison of synthetic MRI volumes generated by different models. The displayed images highlight structural variations in the liver and vascular structures. Compared to other models, 3D-LLDM generates more anatomically consistent images, accurately capturing finer details in liver margin sharpness and vascular structure delineation. This qualitative analysis makes 3D-LLDM well-suited for data augmentation in MR imaging tasks.

4.3. Evaluation of Synthetic MR Volumes in Downstream Tasks

To evaluate the impact of 3D-LLDM-generated synthetic MR data on downstream tasks, we conducted comprehensive 3D segmentation experiments using real and synthetic label-volume pairs. We compared two training configurations: models trained exclusively on real data versus models trained on real data augmented with an equal proportion of synthetic samples.

The downstream 3D segmentation tasks encompassed four distinct scenarios: (1) hepatocellular carcinoma (HCC) tumor segmentation, (2) venous segmentation targeting portal and hepatic veins, (3) liver segmentation treating all anatomical structures except background as liver tissue, and (4) multi-class segmentation that simultaneously delineates portal veins, hepatic veins, and tumors as separate classes.

We evaluated five state-of-the-art CNN-based segmentation architectures: U-Net, ResUNet, WideResUNet, DynUNet, and VNet with increased channel depth. All models were implemented using the MONAI framework [28] and optimized using a combination of Dice loss and cross-entropy loss until validation loss convergence.

CNN Model	Segmentation	R	R + S	Improvement
U-Net	Liver-Only	**0.9650**	**0.9662**	+0.124%
	Vein-Only	0.7905	**0.8667**	+9.640%
	HCC-Only	0.7334	**0.8152**	**+11.153%**
	Multi-Class	0.6968	**0.7014**	+0.660%
ResUNet	Liver-Only	0.9633	**0.9634**	+0.010%
	Vein-Only	0.7197	**0.7902**	+9.796%
	HCC-Only	0.7108	**0.7646**	+7.569%
	Multi-Class	0.6601	**0.6652**	+0.773%
WideResUNet	Liver-Only	**0.9657**	**0.9661**	+0.041%
	Vein-Only	0.7230	**0.7989**	**+10.498%**
	HCC-Only	0.7251	**0.7857**	+8.357%
	Multi-Class	0.6961	**0.7020**	+0.848%
DynUNet	Liver-Only	0.9541	**0.9736**	+2.044%
	Vein-Only	0.6958	**0.7540**	+8.365%
	HCC-Only	0.7002	**0.7674**	+9.597%
	Multi-Class	0.6983	**0.7340**	+5.112%
VNet	Liver-Only	0.9289	**0.9677**	+4.177%
	Vein-Only	0.6908	**0.7212**	+4.401%
	HCC-Only	0.6350	**0.6825**	+7.480%
	Multi-Class	0.5584	**0.5937**	+6.322%
Overall Mean Dice	Liver-Only	0.9544	**0.9674**	+1.362%
	Vein-Only	0.7240	**0.7862**	+8.591%
	HCC-Only	0.7009	**0.7631**	+8.874%
	Multi-Class	0.6619	**0.6793**	+2.629%

Table 2. Comparison of segmentation performance using real data only (R) versus real data augmented with synthetic data (R + S) across different CNN architectures and segmentation tasks. The results demonstrate the effectiveness of synthetic data augmentation, with HCC-Only segmentation showing the highest individual improvement (U-Net: +11.153%) and overall improvement (+8.874% average), followed by Vein-Only segmentation (+8.591% average). **Best** and **2nd-best** results are highlighted.

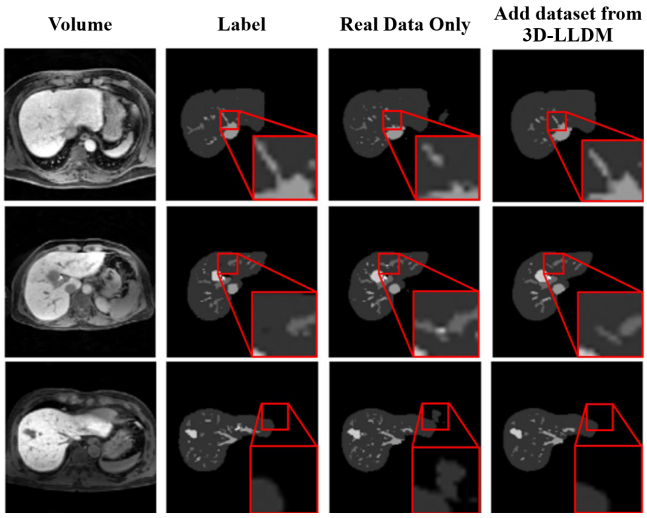

Fig. 4. Qualitative comparison of multi-class segmentation results using U-Net with different training datasets. Each row represents a test sample, while the columns correspond to (1) input MR image, (2) ground truth segmentation, (3) segmentation using only real training data, and (4) segmentation using real+synthetic data from 3D-LLDM. Notably, the 2nd and 4th columns show a more continuous middle hepaticvein, whereas the 3rd column shows a discontinuous appearance.

Training employed the Adam optimizer with a learning rate of 10^{-4} and momentum of 0.95, requiring up to 18 hours per model on an NVIDIA A100 GPU.

Table 2 presents the segmentation performance comparison across different CNN architectures and training configurations. The incorporation of synthetic 3D-LLDM data consistently improved segmentation performance across most experimental settings. For liver segmentation, the average Dice score across all models reached 0.9674 with synthetic data augmentation, representing a 1.362% improvement over real-data-only training. The most substantial improvements were observed in HCC segmentation, where synthetic data augmentation achieved an average improvement of 8.874%, with U-Net showing the highest individual gain of 11.153%. Vein segmentation also benefited significantly from synthetic data, with an average improvement of 8.591%. These results demonstrate that label-guided synthesis particularly enhances performance for complex anatomical structures such as vessels and tumors, where data scarcity is a common challenge.

Figure 4 provides a qualitative comparison of the multi-class segmentation task with U-Net using different training datasets. In the fourth column, the segmentation model trained with synthetic data demonstrates improved accuracy, particularly in segmenting the middle hepatic vein, closely aligning with the ground truth labels shown in the second column. These findings validate that augmenting the training dataset with synthetic samples from 3D-LLDM enhances segmentation performance in downstream tasks.

5. DISCUSSION AND CONCLUSION

This study introduced the 3D-LLDM designed to generate high-resolution synthetic MR volumes along with the corresponding segmentation labels. Leveraging ControlNet for spatial guidance, our model ensured anatomically consistent synthesis and outperformed conventional generative models, achieving the lowest FID score and improving segmentation performance for the liver and multi-class tasks. Our study demonstrates the capability of 3D-LLDM to mitigate data scarcity in medical imaging, most notably in MR imaging. Subsequent work will aim to expand this approach across diverse imaging modalities and to exploit domain adaptation for improved generalization.

6. REFERENCES

[1] Lei Cai, Jingyang Gao, and Di Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Annals of translational medicine*, vol. 8, no. 11, pp. 713, 2020.

[2] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[3] Xiaodong Liu Hao Wu, Qi Liu, "A review on deep learning approaches to image classification and object segmentation," *Computers, Materials & Continua*, vol. 60, no. 2, pp. 575–597, 2019.

[4] Xiaoqing Liu, Kunlun Gao, Bo Liu, Chengwei Pan, Kongming Liang, Lifeng Yan, Jiechao Ma, Fujin He, Shu Zhang, Siyuan Pan, and Yizhou Yu, "Advances in deep learning-based medical image analysis," *Health Data Science*, vol. 2021, pp. 8786793, 2021.

[5] Anamika Jha and Hitoshi Iima, "Ct to mri image translation using cyclegan: A deep learning approach for cross-modality medical imaging.," in *ICAART (3)*, 2024, pp. 951–957.

[6] Seung Kwan Kang, Hyun Joon An, Hyeongmin Jin, Jungin Kim, Eui Kyu Chie, Jong Min Park, and Jae Sung Lee, "Synthetic ct generation from weakly paired mr images using cycle-consistent gan for mr-guided radiotherapy," *Biomedical engineering letters*, vol. 11, no. 3, pp. 263–271, 2021.

[7] YiRang Shin, Jaemoon Yang, and Young Han Lee, "Deep generative adversarial networks: Applications in musculoskeletal imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 3, pp. e200157, 2021.

[8] 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)*. IEEE, 2023, pp. 1–5.

[9] 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*, 2023, pp. 21152–21164.

[10] Yi Zhou, Xiaodong He, Shanshan Cui, Fan Zhu, Li Liu, and Ling Shao, "High-resolution diabetic retinopathy image synthesis manipulated by grading and lesions," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*. 2019, vol. 11766 of *Lecture Notes in Computer Science*, pp. 505–513, Springer.

[11] Lvmin Zhang, Anyi Rao, and Maneesh 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.

[12] Zhihan Ju, Wanting Zhou, Longteng Kong, Yu Chen, Yi Li, Zhenan Sun, and Caifeng Shan, "Hagan: Hybrid augmented generative adversarial network for medical image synthesis," *arXiv preprint arXiv:2405.04902*, 2024.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[15] Alex Ling Yu Hung, Kai Zhao, Haoxin Zheng, Ran Yan, Steven S Raman, Demetri Terzopoulos, and Kyunghyun Sung, "Med-cdiff: Conditional medical image generation with diffusion models," *Bioengineering*, vol. 10, no. 11, pp. 1258, 2023.

[16] Yaqing Shi, Abudukelimu Abulizi, Hao Wang, Ke Feng, Nihemaiti Abudukelimu, Youli Su, and Halidanmu Abudukelimu, "Diffusion models for medical image computing: A survey," *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 357–383, 2025.

[17] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical image analysis*, vol. 88, pp. 102846, 2023.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[19] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah, "Diffusion models in vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[20] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al., "Generatect: Text-conditional generation of 3d chest ct volumes," in *European Conference on Computer Vision*. Springer, 2024, pp. 126–143.

[21] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou, "Synthetic tumors make ai segment tumors better," *arXiv preprint arXiv:2210.14845*, 2022.

[22] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou, "From pixel to cancer: Cellular automata in computed tomography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 36–46.

[23] 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*, vol. 41, no. 9, pp. 2510–2520, 2022.

[24] Linshan Wu, Jiaxin Zhuang, Xuefeng Ni, and Hao Chen, "Freetumor: Advance tumor segmentation via large-scale tumor synthesis," *arXiv preprint arXiv:2406.01264*, 2024.

[25] 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*, vol. 27, no. 9, pp. 4373–4384, 2023.

[26] 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*, 2022, pp. 10684–10695.

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

[28] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al., "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[30] Sihong Chen, Kai Ma, and Yefeng Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.