Synthetic Multimodal Data Generation and Training Optimization For Computed Tomography Cardiac Imaging Applications

Masaki Ikuta*
GE HealthCare

Quanqi Hu
Texas A&M University

Ashok Vardhan Addala
GE HealthCare

Ravi Soni
GE HealthCare

Gopal Avinash
GE HealthCare

Parminder Bhatia
GE HealthCare

Abstract

Computed Tomography (CT) cardiac imaging is among the most complex visualization techniques within CT organ imaging procedures, primarily due to the dynamic nature of human hearts that are constantly working and pumping blood. To accurately capture the organs, CT scanners must perform fast scans to produce a "snapshot" of a human heart, yet their temporal resolution remains limited by CT systems' mechanical constraints. Recently, Generative AI has gained considerable attention, with extensive research exploring its potential to generate detailed synthetic images. In medical imaging, these techniques potentially offer a promising solution to the scarcity of CT cardiac data stemming from the aforementioned challenges. While these synthetic images appear highly realistic, an important question arises: Can they effectively support downstream tasks, such as semantic image segmentation? In this paper, we introduce a novel latent diffusion model as a generative model for 3D CT cardiac imaging, capable of producing multi-modal data including synthetic CT cardiac images alongside corresponding heart substructures. These multi-modal synthetic data are utilized in both the pre-training phase (via Self-Supervised Learning) and the fine-tuning phase (via Supervised Learning). Through extensive experimentation, we demonstrate that the synthetic data generated by our generative model significantly enhances 3D CT cardiac image segmentation performance, contributing to more accurate and robust diagnoses.

1 Introduction

Computed Tomography (CT) imaging has celebrated its 50th anniversary recently. The technology is still rapidly evolving as of this writing in 2024. With this medical imaging technique, clinicians can have detailed visualization of the internal structures of human bodies including bones, organs, blood vessels, and soft tissues (Hsieh, 2009; Buzug, 2008; Ikuta and Zhang, 2023b). CT imaging can be used to take scans of many human organs. CT cardiac imaging remains one of the most challenging visualization techniques among numerous CT organ imaging procedures. This is because of the dynamic nature of human hearts, constantly moving and pumping blood (Hsieh, 2009; Buzug, 2008; Ikuta and Zhang, 2022). Organ segmentation is a critical task in medical imaging. CT cardiac image segmentation is to specify and classify different structures and parts of human hearts. CT cardiac chamber image segmentation is one of the most challenging tasks in medical image semantic segmentation tasks due to the complex anatomy of the human heart, variability in heart size among

*The corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

patients, difficulties with temporal resolution, and the dynamic motions of the heart, among other factors.

Recently, there has been a substantial advancement in the use of Deep Learning (DL) techniques for medical image segmentation. The performance of these models largely hinges on access to large, high-quality annotated datasets (Ikuta and Zhang, 2023a; Hosseinzadeh Taher et al., 2021). However, obtaining such datasets, especially for 3D CT cardiac image segmentation, is often expensive and time-consuming due to the inherent challenges involved in each image annotation process. A promising way to overcome the shortage of annotated data in CT cardiac imaging is the Self-Supervised Learning (SSL) approach (Hosseinzadeh Taher et al., 2023, 2021), which has achieved tremendous success in fields like Natural Language Processing (NLP) (Ray, 2023; Liu et al., 2023) and Computer Vision (CV) (Chen et al., 2020b; Grill et al., 2020; Misra and Maaten, 2020). SSL techniques seek to derive general representations from unlabeled data, which can then be fine-tuned for various tasks, even when labeled data is scarce (Haghighi et al., 2021). Despite the growing number of self-supervised algorithms in medical imaging (Azizi et al., 2023; Haghighi et al., 2020; Hosseinzadeh Taher et al., 2022), existing SSL methods struggle to capture meaningful representations from 3D CT cardiac image volumes due to the lack of consideration for the dynamic nature of the human heart in the design of their pre-text tasks.

The latent Diffusion Model (LDM) (Ho et al., 2022; Rombach et al., 2021) is a type of generative model used in DL. It uses the concept of a diffusion process (Ho et al., 2022) to generate new image data. This diffusion process is an image generation technique developed based on a stochastic process that describes how data changes over time. It gradually converts from a simple probability distribution such as Gaussian noise to an image (or an image volume if it is three-dimensional). While a conventional diffusion process is performed on the input image space, the LDM performs the diffusion process in a latent space. There are a couple of advantages of using the LDM. The first advantage is GPU (Graphical Processing Unit) memory efficiency. By artificially introducing and removing noises in the latent space, we can reduce the GPU memory consumption required for training and validation. This leads to faster training and validation or enables to use a larger image matrix size. The second advantage is image quality. We can generate visually striking image samples from complex data distributions, especially those found in medical imaging. In its training process, the LDM learns how to reverse the diffusion process. In other words, it learns how to gradually recover original images from an artificially added Gaussian noise in the latent space. Once training is completed, a trained LDM can start generating images. Images can be generated by converting random noise in the latent space into samples of the learned data distribution through the learned de-noising process. LDMs are currently used in many computer vision and medical image processing applications, such as image synthesis, image restoration, and super-resolution (Ho et al., 2022; Rombach et al., 2021). However, they have not yet been extensively used in CT cardiac imaging applications.

In this paper, we introduce a novel Latent Diffusion Model (LDM) as a generative model for CT cardiac imaging, designed to produce multi-modal data, including synthetic CT cardiac images along with their corresponding cardiac sub-structures. These generated sub-structures are transformed into segmentation labels and used to train a semantic image segmentation model. Given that CT cardiac imaging is inherently three-dimensional, our LDM is developed as a 3D image generator. The generative model learns complex data distributions from ground truth data, and the generated synthetic data are combined with ground truth data to enhance the size of the training and validation datasets. This approach offers a promising solution to the scarcity of CT cardiac data caused by the inherent challenges of CT cardiac scans, ultimately contributing to more accurate and robust diagnoses.

In summary, the main contributions of this work are:

- We present a new 3D Latent Diffusion Model (LDM) as a generative model for CT cardiac imaging. The model generates synthetic CT cardiac images as well as corresponding human heart sub-structures. The sub-structure data are converted to segmentation labels. Generated data are used in a cardiac chamber image semantic segmentation to enhance training and validation data.
- In addition, we propose a new Self-Supervised Learning (SSL) training framework with the LDM, where we have distinct data augmentation techniques to enhance the variety of diffusion-generated data. The generated images are used along with Masked Image

Modeling (MIM) as part of SSL to help our semantic segmentation model learn the visual representation of 3D cardiac image volumes more efficiently.

- Furthermore, we conduct qualitative image analysis on diffusion-generated images and segmentation labels and scrutinize the data for positives and negatives in terms of CT cardiac imaging.
- Finally, we conduct extensive experiments to verify the superior performance of our proposed framework.

2 Related Work

In this section, we discuss related work and relevant topics to our research.

CT Cardiac Image Segmentation is a challenging problem. Although deep learning techniques have been widely applied to cardiac image segmentation in MRI and ultrasound (Chen et al., 2020a), there has been comparatively little research focused on CT images. Dormer et al. (2018) used a 2D CNN model to segment four heart chambers from patches extracted from 3D CT scans. Other methods (Tong et al., 2018; Wang and Smedby, 2018) have integrated a 3D fully convolutional network (FCN) with a localization network to first detect the region of interest for whole heart segmentation in multi-modal settings. Morris et al. (2020) proposed a 3D U-Net-based design with multiple enhancements to segment cardiac substructures in MRI and CT pairs, while Harms et al. (2021) developed a segmentation network leveraging regional convolutional neural networks. Wang et al. (2022) introduced a hybrid model that combines CNNs and transformers for cardiac segmentation, and Momin et al. (2022) designed a method using mutually enhancing networks to localize and segment cardiac substructures simultaneously in a bootstrapping manner. A common issue across these studies is the limited availability of annotated data for training deep models in cardiac chamber segmentation. Unlike previous work, our approach addresses this challenge by introducing a self-supervised learning method with a latent diffusion model for 3D cardiac CT image segmentation.

Self-Supervised Learning (SSL) is a promising approach. Given the limited availability of large-scale annotated datasets, as discussed in the previous section, the SSL holds significant assurance for medical imaging applications. In this framework, a neural network is trained on a carefully designed pre-text task using unlabeled data, and the learned representations are later fine-tuned for specific tasks with annotated data (Haghighi et al., 2021; Hosseinzadeh Taher et al., 2021). State-of-the-art SSL approaches can be roughly divided into two categories: Instance Discrimination Learning (IDL) and Masked Image Modeling (MIM). Instance discrimination methods (He et al., 2020; Azizi et al., 2023; Chen et al., 2020c; Chaitanya et al., 2020; Haghighi et al., 2023) treat each image as a unique class and aim to learn image representations that are robust to image distortions. In contrast, MIM methods (Xie et al., 2022; He et al., 2022; Zhou et al., 2021) mask random regions of an image and train a model to predict the masked areas. Unlike these existing SSL techniques, we introduce an SSL approach using a latent diffusion model, where the diffusion process learns the data distribution of GT images and generates new synthetic data. Self-supervised learning is then applied using the synthetic data, enabling the model to acquire general knowledge from a larger pool of generated images. This process provides valuable contextual information for tackling more complex tasks and results in more generalizable features for cardiac CT imaging.

Latent Diffusion Model is an active research area in recent years. Ho et al. (2022) proposed a novel high-quality image synthesis technique using diffusion probabilistic models. Their diffusion process is conducted in the image space. The target applications are computer vision sample generations. Therefore, the method is limited to 2D image generation. Rombach et al. (2021) proposed a new high-quality image synthesis technique using diffusion models. Their method is to conduct the diffusion process in the latent space rather than in the image space for computational efficiency. It turns out that the latent diffusion model can create more striking image quality than the ones using the image space. This technique again targets computer vision applications. Therefore, the method is limited to 2D image generation as well. Txurio et al. (2023) applied the latent diffusion models to CT imaging applications. While the method is proven effective in generating high-quality CT images, it is limited to 2D image generation. Nor the method cannot create segmentation labels. Khader et al. (2023) proposed a new latent diffusion model for CT imaging applications where their method creates 3D imaging volumes. These images are used in their self-supervised learning (SSL) to increase their

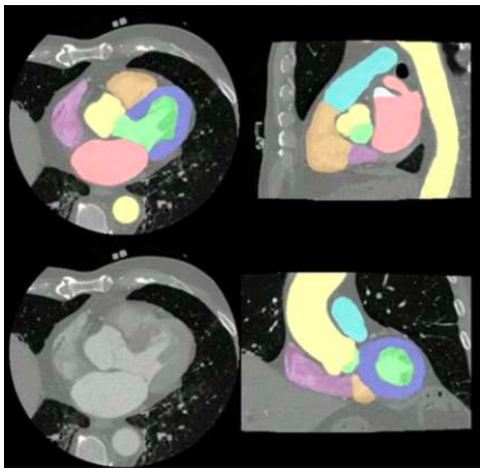

Figure 1: An example of a CT cardiac chamber image annotated by clinical experts. Each unique color represents different heart substructures, including the left atrium (LA), left ventricle (LV), right atrium (RA), right ventricle (RV), myocardium (MYO), aorta (AO), pulmonary artery (PA), and left atrial appendage (LAA).

image segmentation performance. However, their method only creates images and is not scalable to create segmentation labels. In this paper, we propose a new latent diffusion model for CT cardiac imaging where the model produces both images and labels. It learns a complex data distribution from ground truth images and labels. These generated data are supplement to ground truth data to boost the amount of training and validation data. CT cardiac imaging is three-dimensional in nature, thus, we create the latent diffusion model as a 3D image generator.

3 Method

In this section, we discuss our data preparation followed by our method in three parts, synthetic data generation by the latent diffusion model, self-supervised learning as a pre-training, and finally our fine-tuning process for the 3D image segmentation.

3.1 Data preparation

In this research, we use one of our proprietary data sets of 3D CT cardiac imaging. The data set has been collected from 32 different hospitals in 10 different countries worldwide. In the data set, there are eight heart substructures, that are left atrium (LA), left ventricle (LV), right atrium (RA), right ventricle (RV), myocardium (MYO), aorta (AO), pulmonary artery (PA), and left atrial appendage, (LAA) which were manually annotated by clinical experts on 262 cardiac CT Angiography series. The total number of patients is 262. The total number of images in the data set is 65418. The size of each 3D image volume is 512x512 matrix size with different numbers of images in the z-direction. The z size varies from a minimum of 140 to a maximum of 560 where the median number of images is 224. Each image volume is normalized to [0, 1] by -1000 Hounsfield Unit (HU) and +2000 HU. Among them, 168 series were used for training, and 43 series were used as the validation data set for saving the best checkpoint models. In addition, a separate, fully annotated set of 51 cases served as an independent test data set for a model evaluation. Furthermore, we resize 512x512 images and masks to 256x256 to reduce memory footprints. Regarding lables, the original labels have 8 channels.

3.2 Latent Diffusion Model

We consider the well-known latent diffusion model (LDM) (Rombach et al., 2021) for the data generation due to its efficiency in terms of computational resources and high quality of generative images. The training of LDM has two phases. First, we train an autoencoder to encode the original input images onto a lower-dimensional representative space, which is a latent representation of the pixel space. Then, we train a diffusion model on the learned latent space. As a result, LDM is much

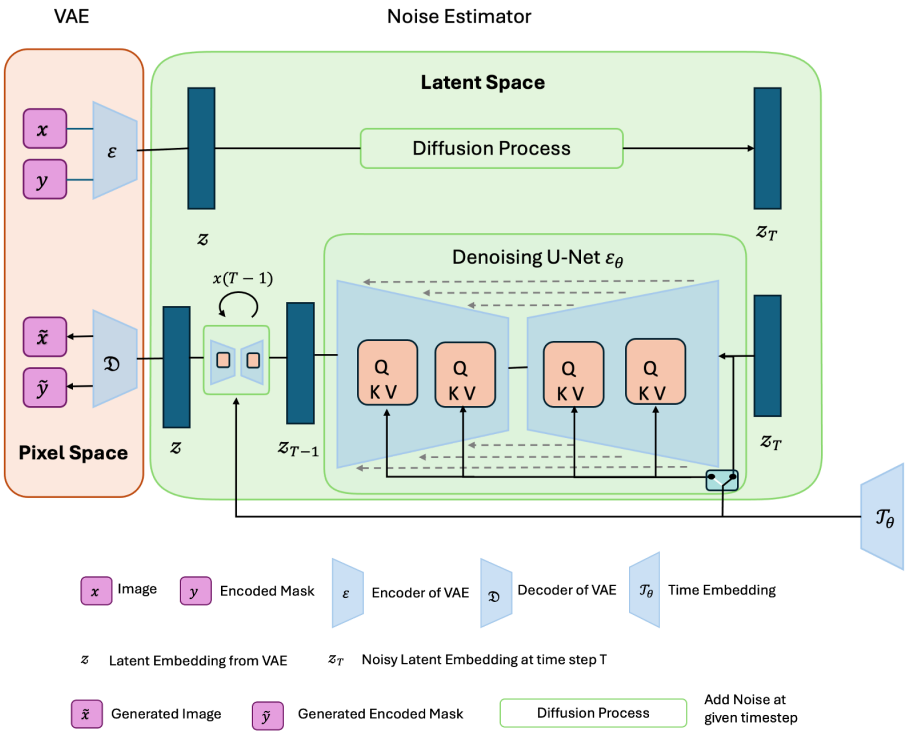

Figure 2: Three Dimensional Latent Diffusion Model Training and Image Generation Process for 3D CT Cardiac Chamber Image Segmentation.

more efficient than training diffusion models directly on the pixel space. However, the original LDM we use is designed for generating 2D images. To adapt it to 3D image volumes, we add the depth dimension to the model so that, during the training process, the output of the encoder has the shape of (width=256, length=256, depth=Z, channels=3) where Z is the number of images in the z-direction and we set it to 160 in our case. The original number of channels in segmentation labels is eight. To generate synthetic human cardiac sub-structures that are later converted to segmentation labels, we encode eight channels into three channels (RGB) without compromising the VAE performance. This is also helpful in reducing the memory footprint and GPU memory usage. These dimension data are encoded into the latent space, and they are the input to the diffusion model.

Moreover, to better assist the segmentation training, we modify the architecture of the autoencoder so that the LDM can generate a 3D image volume with its corresponding human cardiac sub-structures (that are later converted to segmentation mask volumes) simultaneously. Figure 2 shows our three-dimensional latent diffusion model architecture for the 3D CT cardiac chamber image segmentation. During training, we first convert the image x_i and the mask x_m into vectors $\tilde{x}_i, \tilde{x}_m$ of the same shape by two one-layer encoders (i.e., ε_i and ε_m) separately. Then we take the summation of the encoded image and mask and encode it onto the latent space, i.e., $z = \varepsilon_0(\tilde{x}_i + \tilde{x}_m)$. As part of the architecture design, we also explore both an addition and a concatenation to get the unified z vector, however, both produce the same level of final reconstruction performances in the 3D VAE. Thus, we choose the addition over the concatenation for, again, GPU memory efficiency.

3.3 Self-Supervised Learning as a Pre-training

Given generative images and masks sampled from the LDM, we conduct self-supervised learning to pre-train our model before fine-tuning it for the 3D image segmentation task. Recently, Taher et al. (2023) has shown that the SSL pretraining on ground truth images can greatly improve the segmentation performance on 3D cardiac CT images. Moreover, Khader et al. (2023) mentioned that the SSL on synthetic images can also improve segmentation performance on 3D medical images in

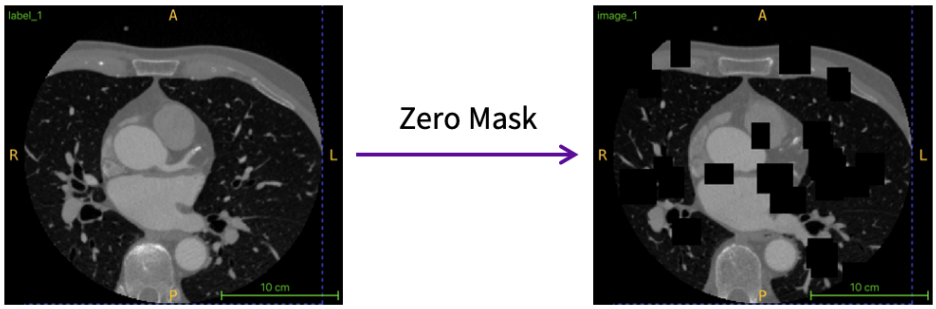

Figure 3: Zero masking process in Self-Supervised Learning (pre-train).

general. However, since there is no implementation nor results for CT images available in Khader et al. (2023), we are motivated to explore a better way to leverage the generative data for SSL pretraining and segmentation finetuning on 3D cardiac CT images.

For the SSL pretraining, following the work Taher et al. (2023), we first mask a portion of the original image with zeros, and then we train the model to reconstruct the original image. The zero masking process is illustrated in Figure 3. The model is trained by minimizing the L2 norm of pixel value difference between the original image and the reconstructed image, i.e.,

$$L_{SSL} = \mathbb{E}_{x \sim X} \|x - f(\tilde{x})\|_2 \tag{1}$$

where x is an image from the dataset X, $\tilde{x}$ is the image x with zero masks, and f is the model we aim to train.

Following the method Khader et al. (2023), we conduct the self-supervised learning with diffusion-generated data. Specifically, we employ a 3D U-Net (Ronneberger et al., 2015) as the primary architecture of our proxy model; nevertheless, alternative architectures, such as vision transformers (Tang et al., 2022), can also be used seamlessly. We mask out 25 blocks with a probability of 0.8. We utilize the minimum of 8x8 pixels, and the maximum of 16x16 pixels for the block's spatial sizes. The masking block sizes and locations are randomly selected. We use the AdamW optimizer with a learning rate of 0.001. We use the early-stopping technique with a patience of 50 using 10% of training data as the validation set. We save the best model based on the validation loss and transfer the best model to the target task.

3.4 Three Dimensional Image Segmentation as a Fine-tuning process

In the fine-tuning phase (the target task), we mix the diffusion-generated data with the ground truth data and train the segmentation model, where we keep the encoder weights and randomize the de-coder weights from the pre-training phase. In this phase, all the downstream model's parameters are fine-tuned. This mixed dataset is distinct from the state-of-the-art method (Khader et al., 2023), where they use the ground truth dataset in the fine-tuning phase. Our fine-tuning with the mixed dataset only becomes possible because our latent diffusion model generates both image volumes and segmentation masks while their method (Khader et al., 2023) only produces image volumes. We again use the AdamW with a learning rate of 0.001. To prevent over-fitting, we employ an early-stopping technique with a patience of 10 using 10% of the training data as the validation set. We evaluate the segmentation performance using the Dice coefficient.

4 Experiments

In this section, we present our experimental results, where we show some example images and labels from our latent diffusion model followed by example results of our data augmentation strategies, some observations on them, and finally our quantitative results compared to our baseline methods.

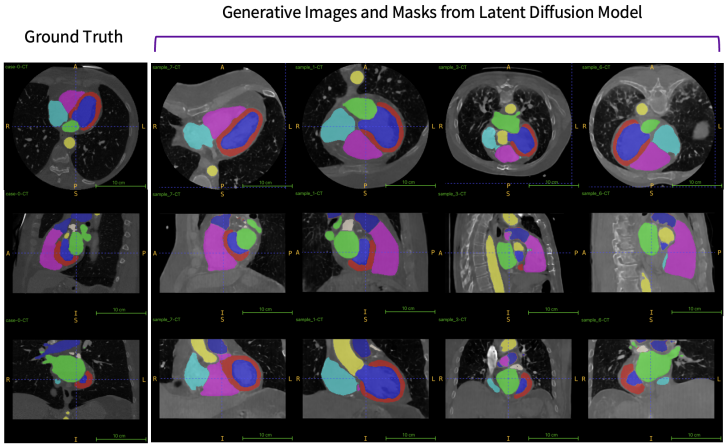

Figure 4: Examples of Ground Truth (GT) images, diffusion-generated images, and the corresponding segmentation labels. The GT images and diffusion-generated images are not necessarily pairs in this figure. There are three different GT image examples from different patients and different anatomy locations. We pick four different diffusion-generated images from similar locations for each GT example.

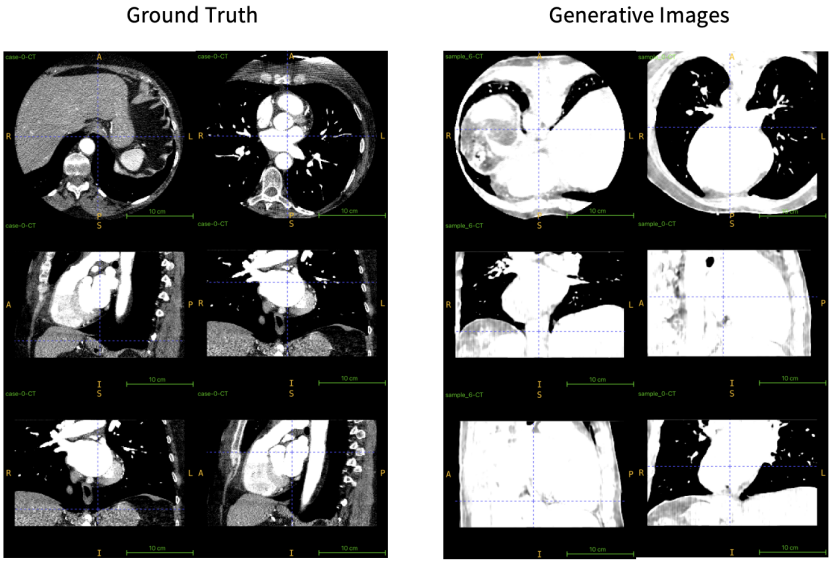

Figure 5: Data augmentation: generative images with shifted contrast under soft tissue view.

First of all, figure 4 shows examples of diffusion-generated images and the corresponding segmentation labels. As we can see, these generative images are striking, and they signify the promising capability of the latent diffusion model.

Second, in the latent diffusion model, we use some data augmentation techniques to increase the varieties of the dataset so that the model can produce more diverse data rather than creating replicas of the GT data. We perform a center crop on each image volume and resize them back to the original sizes in the xy-axis. The location of the center crop is randomized. In addition, we conduct a horizontal and vertical flip on input data with a probability of 20%. Furthermore, we boost pixel values by about 50 Hounsfield Units (HU) with a probability of 5%. Figure 5 shows examples of shifted contrast. The left-hand side of the figure shows some ground truth images. The right-hand side shows diffusion-generated images. They are the results of our contrast boosts. These contrast boosts only happen on 5% of generative images because the data augmentation is used with 5% probability. They are great additions to our dataset for the following reasons. In real clinical settings,

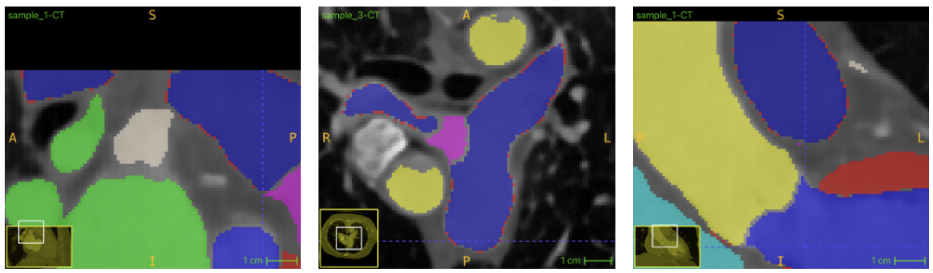

Figure 6: In the generative masks, we observe boundary mislabeling issues

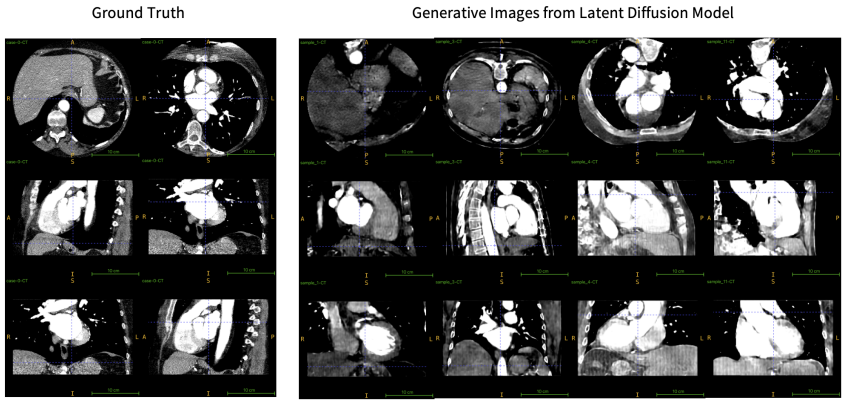

Figure 7: Example of smoother soft tissues on diffusion-generated images

clinicians often use contrast agents (chemical liquid injected into blood vessels) to create contrasts in blood flows from surrounding soft tissues. This makes it easier for clinicians to visualize blood flow in cardiac chambers. However, some patients are allergic to these chemical agents, and clinicians choose not to inject the chemical liquid. As a result, blood vessels in these patients do not have contrast boosts. Therefore, it is hard to visualize their blood vessels. Usually, many cardiac images have contrast boosts, but some of them do not have the boosts. Thus, using such a data augmentation technique will help increase the variety of diffusion-generated cardiac datasets.

Furthermore, figure 6 shows some minor problems on the diffusion-generated segmentation labels. As we can see, some boundary pixels on segmentation labels are not necessarily cut and clean. For example, the boundary pixels of the purple (or blue) segment on the center image in figure 6 have many red dots. We believe this may come from the fact that there are always gaps among segmentation labels on the GT data, and our latent diffusion model might get confused about how to segment boundary pixels. The examples in figure 6 show many red dots, however, we saw the problem is not limited to the red label. This problem can happen with any segmentation labels. While we do not believe this problem influences our fine-tuning segmentation performance, we want to make a note of this phenomenon in this article.

Moreover, there are some more interesting observations on our diffusion-generated images, particularly on soft tissues. In practice, the latent diffusion model seems to be struggling to remove how much Gaussian noise it should remove from images. The original GT images already have some Gaussian noises due to equipment electronic noise at the time of patient scanning. The model needs to keep these Gaussian noises and should only remove the Gaussian noises we add as part of the diffusion process. However, the model seems struggling with noise removal operations. Of course, we add noise not on images themselves, but on the latent embeddings. However, added Gaussian noises on the latent space seem to influence "look and feel" on image space, and the latent diffusion model tends to get confused with Gaussian noises that have different origins. As a result, diffusion-generated images tend to have smoother surfaces on soft tissues on images. Figure 7 shows some examples of such a problem. On the left-hand side, some GT examples are shown where soft tissues have realistic

Table 1: Segmentation model performance from

Method	Average	AO	LAA	LA	LV	MYO	PA	RA	RV
Baseline	0.7743	0.9275	0.5790	0.8061	**0.8298**	0.8026	0.6888	0.7705	0.7902
Medical Diffusion	0.7855	0.9365	0.6093	0.8061	0.8278	**0.8141**	0.6976	0.7850	0.8078
Ours	**0.7908**	**0.9396**	**0.6109**	**0.8288**	0.8190	0.8094	**0.7069**	**0.8021**	**0.8100**

noise characteristics. In contrast, the right-hand side shows diffusion-generated images where soft tissue pixels have smoother surfaces, and they are not necessarily realistic look-and-feel. While it is hard to know how this phenomenon influences our fine-tuning phase, we want to look into this problem in the near future.

Finally, we conduct a few experiments to demonstrate the effectiveness of our method. The experimental results are shown in Table 1. First, we trained our network with only GT data. We set the maximum epochs to be 300. However, our early-stopping criteria met at the 25th epoch, and the best model was picked from the 15th epoch, where the patience of the early-stopping was set to 10 epochs. We use this result as our baseline performance. Next, we try to reproduce the state-of-the-art method called Medical Diffusion (Khader et al., 2023). Because their diffusion implementation is not available in the public domain, we use our latent diffusion model to generate images. Our diffusion model does create the corresponding segmentation labels as well. However, medical diffusion is limited to producing only images. Thus, we throw away segmentation labels to reproduce their method. In the fine-tuning process, we only use the ground truth data for the medical diffusion. In this fine-tuning phase, the training is done with the 28th epoch, and the best model is picked from the 18th epoch, where the patience of the early stopping is again set to 10 epochs. We use this result as our state-of-the-art baseline performance. Now, regarding our method, we pre-train our model with diffusion-generated data as the SSL. In the fine-tuning phase, we use the mixed dataset, including diffusion-generated images, and diffusion-generated segmentation labels as well as the GT data. The number of image volumes of the ground truth is 167. We generated three times more data by the latent diffusion model. Therefore, the total number of image volumes is 668. In the fine-tuning phase, the training is done with the 35th epoch, and the best model is picked from the 25th epoch, where the patience of the early stopping is again set to 10 epochs.

Table 1 shows our quantitative results compared to our baseline methods. First, the medical diffusion (Khader et al. (2023), the second row in the table) produces a nice improvement of DICE 1.12% over the baseline (the first row, trained from scratch). The standard deviation of the baseline method is about 0.27%. Therefore, the medical diffusion has a statistically significant improvement on the average DICE score. This signifies the effectiveness of the pre-train phase with diffusion-generated data. Now, regarding our method, there is again a meaningful improvement on the average DICE score over the medical diffusion (Khader et al., 2023), where the average DICE score improvement over the medical diffusion is 0.53%. Thus, this is also a statistically significant difference. This improvement comes from the fact that our diffusion-generated data (both image volumes as well as segmentation labels) are used along with the GT data (we call them "mixed dataset") in the fine-tuning phase. This indicates the latent diffusion model does create meaningful segmentation labels in addition to images over the GT data. One interesting observation is that the primary improvement of our method over the medical diffusion comes from two cardiac chambers, which are the Left Atrium (LA) and the Right Atrium (RA). They are two of the four biggest cardiac chambers, and they are relatively easy to spot in images. On the other hand, our method does not necessarily improve the segmentation performance on minor cardiac chambers such as Pulmonary Artery (PA) and Left Atrial Appendage (LAA). This is in contrast to what we have hoped, and we want to work on this area in the future.

5 Conclusion

In this paper, we present a novel latent diffusion model for CT cardiac imaging, where the generative model produces multi-modal data, including synthetic CT cardiac images and their corresponding cardiac sub-structures. The model is trained using diverse data augmentation techniques to increase the variety of the generated data. These synthetic data are utilized in both the pre-training and the fine-tuning phase. During fine-tuning, we use three times more synthetic data than ground truth (GT) data, resulting in the training dataset being four times the size of the original GT dataset. We also perform a qualitative analysis of the diffusion-generated images and segmentation labels, discussing

both the strengths and limitations in the context of CT cardiac imaging. Our approach improves the final segmentation performance, achieving a 1.65% average increase in DICE score over the baseline (trained from scratch) and a 0.53% improvement over the current state-of-the-art medical diffusion method Khader et al. (2023). This technique provides a promising solution to the scarcity of CT cardiac data due to the challenges of cardiac imaging, leading to more accurate and robust diagnoses. Looking ahead, we plan to extend this 3D latent diffusion model to other CT organ datasets and explore its application across other medical imaging modalities such as MRI, X-ray, and ultrasound.

PII and IRB Information

This study is conducted in accordance with ethical standards. Due to the nature of the study, which uses the existing data set mentioned above and does not contain Personal Identifiable Information (PII), no Institutional Review Board (IRB) review is required.

References

S. Azizi, L. Culp, J. Freyberg, and et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7:756–779, 2023.

T. Buzug. *Computed tomography: from photon statistics to modern cone-beam ct*. Springer, 2008.

K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/949686ecef4ee20a62d16b4a2d7ccca3-Paper.pdf.

C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7, 2020a. ISSN 2297-055X. doi: 10.3389/fcvm.2020.00025.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020b.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020c. URL https://proceedings.mlr.press/v119/chen20j.html.

J. D. Dormer, L. Ma, M. Halicek, C. M. Reilly, E. Schreibmann, and B. Fei. Heart chamber segmentation from CT using convolutional neural networks. In B. Gimi and A. Krol, editors, *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, page 105782S. International Society for Optics and Photonics, SPIE, 2018. doi: 10.1117/12.2293554. URL https://doi.org/10.1117/12.2293554.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

F. Haghighi, M. R. Hosseinzadeh Taher, Z. Zhou, M. B. Gotway, and J. Liang. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 137–147, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8.

F. Haghighi, M. R. H. Taher, Z. Zhou, M. B. Gotway, and J. Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 40(10):2857–2868, 2021. doi: 10.1109/TMI.2021.3060634.

F. Haghighi, soumitra ghosh, S. Chu, H. Ngu, M. Hejrati, H. H. Lin, B. Bingol, and S. Hashemifar. Self-supervised learning for segmentation and quantification of dopamine neurons in parkinson's disease. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

J. Harms, Y. Lei, S. Tian, N. S. McCall, K. A. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang. Automatic delineation of cardiac substructures using a region-based fully convolutional network. *Medical Physics*, 48(6):2867–2876, 2021. doi: https://doi.org/10.1002/mp.14810. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14810.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS 2020*, 2022. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

M. R. Hosseinzadeh Taher, F. Haghighi, R. Feng, M. B. Gotway, and J. Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87722-4.

M. R. Hosseinzadeh Taher, F. Haghighi, M. B. Gotway, and J. Liang. Caid: Context-aware instance discrimination for self-supervised learning in medical imaging. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 535–551. PMLR, 06–08 Jul 2022.

M. R. Hosseinzadeh Taher, M. B. Gotway, and J. Liang. Towards foundation models learned from anatomy in medical imaging via self-supervision. *arXiv:2309.15358*, 2023. URL https://arxiv.org/abs/2309.15358.

J. Hsieh. *Computed tomography: principles, design, artifacts, and recent advances*. Wiley, 2009.

M. Ikuta and J. Zhang. A deep recurrent neural network with fista optimization for ct metal artifact reduction. *IEEE Transactions on Computational Imaging*, 8:961–971, 2022. doi: 10.1109/TCI.2022.3212825.

M. Ikuta and J. Zhang. A deep convolutional gated recurrent unit for ct image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10612–10625, 2023a. doi: 10.1109/TNNLS.2022.3169569.

M. Ikuta and J. Zhang. TextureWGAN: texture preserving WGAN with multitask regularizer for computed tomography inverse problems. *Journal of Medical Imaging*, 10(2):024003, 2023b. doi: 10.1117/1.JMI.10.2.024003. URL https://doi.org/10.1117/1.JMI.10.2.024003.

F. Khader, G. Mueller-Franzes, S. T. Arasteh, T. Han, C. Haarburger, M. F. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn. Medical diffusion: Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:253384524.

Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017, sep 2023.

I. Misra and L. v. d. Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

S. Momin, Y. Lei, N. S. McCall, J. Zhang, J. Roper, J. Harms, S. Tian, M. S. Lloyd, T. Liu, J. D. Bradley, K. Higgins, and X. Yang. Mutual enhancing learning-based automatic segmentation of ct cardiac substructure. *Physics in Medicine & Biology*, 67(10):105008, may 2022. doi: 10.1088/1361-6560/ac692d. URL https://dx.doi.org/10.1088/1361-6560/ac692d.

E. D. Morris, A. I. Ghanem, M. Dong, M. V. Pantelic, E. M. Walker, and C. K. Glide-Hurst. Cardiac substructure segmentation with deep learning for improved cardiac sparing. *Medical Physics*, 47(2):576–586, 2020. doi: https://doi.org/10.1002/mp.13940. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13940.

P. P. Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023. ISSN 2667-3452. doi: https://doi.org/10.1016/j.iotcps.2023.04.003. URL https://www.sciencedirect.com/science/article/pii/S266734522300024X.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

M. R. H. Taher, M. Ikuta, and R. Soni. Curriculum self-supervised learning for 3d ct cardiac image segmentation. In *ML4H@NeurIPS*, 2023. URL https://api.semanticscholar.org/CorpusID:267760581.

Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740, June 2022.

Q. Tong, M. Ning, W. Si, X. Liao, and J. Qin. 3d deeply-supervised u-net based whole heart segmentation. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 224–232, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75541-0.

M. S. Txurio, K. L.-L. Román, A. Marcos-Carrión, P. Castellote-Huguet, J. M. Santabárbara-Gómez, I. M. Oliver, and M. A. G. Ballester. Diffusion models for realistic ct image generation. In Y.-W. Chen, S. Tanaka, R. J. Howlett, and L. C. Jain, editors, *Innovation in Medicine and Healthcare*, pages 335–344, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-3311-2.

C. Wang and Ö. Smedby. Automatic whole heart segmentation using deep learning and shape context. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 242–249, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75541-0.

J. Wang, S. Wang, W. Liang, N. Zhang, and Y. Zhang. The auto segmentation for cardiac structures using a dual-input deep learning network based on vision saliency and transformer. *Journal of Applied Clinical Medical Physics*, 23(5):e13597, 2022. doi: https://doi.org/10.1002/acm2.13597.

Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, June 2022.

Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models genesis. *Medical Image Analysis*, 67:101840, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101840. URL https://www.sciencedirect.com/science/article/pii/S1361841520302048.