Curriculum Self-Supervised Learning for 3D CT Cardiac Image Segmentation

Mohammad Reza Hosseinzadeh Taher* MHOSSEI2@ASU.EDU
Arizona State University, USA

Masaki Ikuta MASAKI.IKUTA@GE.COM
GE Healthcare, USA

Ravi Soni RAVI.SONI@GE.COM
GE Healthcare, USA

Abstract

Automating the segmentation of various cardiac chamber structures (e.g., pulmonary artery, aorta, etc.) in 3D CT cardiac imaging remains a significant challenge. This challenge primarily arises from the dynamic nature of the human heart and substantial anatomical variations in terms of organ texture, shape, and size across different patients. These factors collectively result in a scarcity of *annotated* data, posing a significant hurdle for training data-hungry deep models. The self-supervised learning (SSL) paradigm offers a promising solution to overcome this obstacle since it eliminates the reliance on massive annotated data for training deep models. However, existing SSL approaches fall short in capturing effective representations from 3D cardiac volumes due to the oversight of the dynamic nature of human hearts in the design of their pretext tasks. To address this challenge, we propose a novel SSL method based on the curriculum learning paradigm, which progressively increases the task difficulty during the pretraining stages. Our method enables the SSL model to initially acquire fundamental knowledge about the data, which can subsequently serve as valuable contextual clues for solving more complex tasks during later stages of pretraining. Our extensive experiments demonstrate that the SSL pre-trained model, trained using our strategy, acquires generalizable representations capable of effectively segmenting various existing cardiac chamber structures.

Keywords: Self-supervised learning, 3D segmentation, CT cardiac imaging

1. Introduction

Accurate segmentation of cardiac chamber structures in 3D CT cardiac imaging—one of the most challenging visualization techniques among numerous CT organ imaging procedures—is crucial for comprehensive cardiac morphology and function analysis, as well as facilitating cardiac arrhythmia detection, cardiac surgery planning, and radiation therapy planning (Chen et al., 2020a; Wang et al., 2022a). The accurate delineation of different cardiac structures can be a very time-intensive process and presents notable challenges, primarily stemming from the dynamic nature of the human heart and substantial anatomical variations in terms of organ texture, shape, and size across different patients. As such, there is a substantial demand for robust automated segmentation systems for 3D CT cardiac imaging in real-world clinical settings.

In recent years, there has been a significant surge in the adoption of deep learning solutions for medical image segmentation tasks. The effectiveness of deep learning models depends heavily on the availability of large-scale, high-quality annotated datasets (Hosseinzadeh Taher et al., 2021). However, acquiring such datasets can be costly and time-consuming, particularly in 3D CT cardiac segmentation applications, where the acquisition of annotated data is inherently difficult due to the aforementioned hurdles. A promising solution for addressing the scarcity of annotated data in CT cardiac imaging is the self-supervised learning (SSL) paradigm (Hosseinzadeh Taher et al., 2023, 2021), which has stunning successes in Natural Language Processing (NLP) (Ray, 2023; Liu et al., 2023) and computer vision applications (Chen et al., 2020b;

* Work done during an internship at GE Healthcare.

Grill et al., 2020; Misra and Maaten, 2020). SSL methods aim to extract general representations directly from unlabeled data. In this paradigm, a neural network is pretrained on a manually crafted (pretext) task, where ground-truth data is available for free. The learned representations can be later fine-tuned on numerous target tasks even with dearth of annotated data (Haghighi et al., 2021). Despite the numerous self-supervised algorithms proposed in medical vision (Azizi et al., 2023; Haghighi et al., 2020; Hosseinzadeh Taher et al., 2022), existing SSL approaches fall short in capturing effective representations from 3D cardiac volumes due to the oversight of the dynamic nature of human hearts in the design of their pretext tasks.

To address this challenge, we propose a novel SSL strategy based on the curriculum learning paradigm. Curriculum learning, inspired by the structured approach of human education, emphasizes the importance of organizing learning examples in a gradual manner rather than randomly. Bengio et al. (2009) pioneered this concept in machine learning, demonstrating its effectiveness in achieving significant generalization. They argued that curriculum learning serves as a specific form of continuation learning method within the realm of machine learning (Bengio et al., 2009; Wang et al., 2022b). Motivated by this, we propose a novel SSL strategy for 3D CT cardiac image segmentation, incorporating the principles of curriculum learning. Considering the complex nature of heart anatomy and substantial anatomical variations across subjects, we initiate the learning process with a simple task. After that, we transfer the acquired knowledge to the next step and progressively increase the next task complexity. This approach allows the SSL model to initially acquire fundamental knowledge about the data, which can subsequently serve as valuable contextual clues for solving more complex tasks during later stages of pretraining. Our extensive experiments demonstrate that this systematic and iteratively adaptive approach leads to acquiring highly generalizable representations tailored to heart anatomy.

In summary, we make the following contributions:

1. A novel self-supervised learning strategy that enhances the segmentation of heart sub-structures in 3D CT cardiac volumes, outperforming the supervised baseline.

2. A set of masking modules tailored for curriculum learning within the SSL paradigm and examining their effectiveness in acquiring generalizable representations.

3. A comprehensive set of experiments that evaluate our proposed SSL learning method across a variety of 8 common heart sub-structures in CT cardiac imaging target tasks.

2. Related work

Cardiac CT Image Segmentation. While deep learning methods have been predominantly proposed for cardiac image segmentation in MRI and ultrasound modalities (Chen et al., 2020a), there has been limited exploration in the context of CT images. Dormer et al. (2018) employed a 2D CNN model to segment four heart chambers using patches extracted from 3D CT images. Other approaches (Tong et al., 2018; Wang and Smedby, 2018) have combined a 3D fully convolutional network (FCN) with a localization network to initially identify the region of interest for whole heart segmentation in multi-modality applications. Morris et al. (2020) introduced a network design based on a 3D U-net with multiple modifications to segment cardiac substructures in paired MRI/CT images. Harms et al. (2021) proposed a segmentation network based on regional convolutional neural networks. Wang et al. (2022a) presented a hybrid model that combines a CNN and a transformer for cardiac segmentation. Momin et al. (2022) introduced a method that utilizes mutual enhancing networks to simultaneously localize and segment each cardiac substructure in a bootstrapping manner. A common challenge across all existing works pertains to the scarcity of annotated data available for training deep models in the domain of heart chamber segmentation. In contrast to prior works, our method seeks to tackle this challenge by developing an effective self-supervised learning approach for Cardiac CT Image Segmentation.

Self-supervised Learning. Due to the limited availability of large-scale annotated datasets, self-supervised learning (SSL) shows great potential for medical applications. In this paradigm, a neural network is trained on a carefully crafted (pretext) task for which ground-truth data is obtained from raw images for free. The learned representations can then be further fine-tuned for various target tasks using only limited annotated data (Haghighi et al., 2021; Hosseinzadeh Taher et al., 2021). The state-of-the-art SSL methods can be broadly cat-

egorized into instance discrimination learning and masked image modeling approaches. 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 separate class and aim to learn representations that remain invariant to image distortions. However, these approaches are sub-optimal for medical applications, which involve images of consistent anatomical structures, and require fine-grained image representations for segmenting small structures (Hosseinzadeh Taher et al., 2022). A line of work incorporated anatomical information to enhance instance discrimination learning for medical tasks (Hosseinzadeh Taher et al., 2023) or added reconstruction (Hosseinzadeh Taher et al., 2022; Tang et al., 2022; Haghighi et al., 2022). On the other hand, masked image modeling (MIM) approaches (Xie et al., 2022; He et al., 2022; Zhou et al., 2021) randomly mask regions within images and train a model to reconstruct the masked portions. Nevertheless, existing MIM-based SSL methods fall short in addressing the distinctive challenges that arise in the context of 3D CT cardiac imaging. For instance, these methods introduce distortions to a significant portion of the images, which can be a challenging task for a model to solve when applied to CT cardiac images due to the dynamic nature of human hearts and the considerable inter-subject and inter-image variations. Consequently, the self-supervised model may struggle to successfully reconstruct the distorted images, leading to a potential failure in capturing rich representations from unlabeled images. In contrast to existing SSL methods, we propose a novel SSL method based on the curriculum learning paradigm, which progressively increases the task difficulty during the pretraining stages. Initially, lightweight distortions are applied to images, allowing the model to learn general knowledge from images and serve as effective contextual clues for solving more complicated tasks during subsequent pretraining stages, resulting in more generalizable features for cardiac CT imaging.

3. Method

Our self-supervised learning strategy, depicted in 1, aims to learn generalized and transferable visual representations. The main intuition behind our learning strategy is the curriculum learning concept: it begins with a simple task and gradually increases the task's complexity during the pre-training procedure.

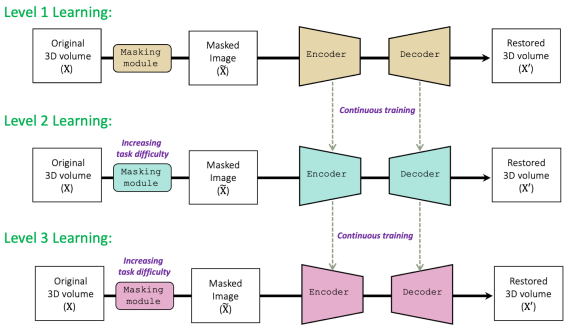

Figure 1: Our proposed Curriculum Self-Supervised Learning framework involves a sequential process. It begins by taking an input, applying masking, and passing it through an encoder-decoder network to restore the original input. As the framework progresses, the difficulty of the tasks gradually increases, and the knowledge acquired at each step is seamlessly transferred to the next stage.

Inspired by this concept, we propose a self-supervised learning (SSL) strategy based on restorative learning.

During pre-training, regardless of their level of complexity, we consistently follow a structured approach. First, we mask a portion of an input image, and then we feed this masked input to our model to reconstruct the original input. Figure 2 illustrates three distinct masking strategies that we employ in our learning strategy:(i) Mean Value Masking: Masks the input using the mean value of designated areas. (ii) Zero Masking: Masks the input by replacing designated parts with zeros, and (iii) Random Noise Masking: Masks the input by introducing random noise to designated portions. Figure 3 illustrates the outputs of our masking module used during our pre-training. Specifically, the leftmost one in Figure 3 corresponds to an output of the masking module at level 1 (representing an easy task), the middle one corresponds to an output at level 2 (representing a medium task), and the rightmost one in Figure 3 relates to an output of the masking module at level 3 (representing a hard task).

For training our SSL model, we start by training the network at level 1, and then the acquired knowledge is passed on to the second level to help the model restore more extensively masked inputs (level 2). Similarly, the knowledge acquired at level 2 is

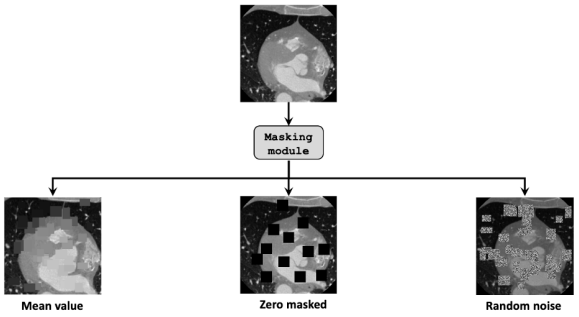

Figure 2: An overview of three masking strategies employed in our learning strategy includes: (i) Mean Value Masking, which masks the input using the mean value of the designated area; (ii) Zero Masking, involving the replacement of designated parts in the input with zeros; and (iii) Random Noise Masking, introducing random noise to the designated portions.

transferred to level 3 to assist the model in restoring highly distorted inputs at level 3. As demonstrated by our experimental results, the knowledge acquired at each level contributes to the next level, enabling the model to extract more effective and fine-grained representations from input data. Our backbone architecture employs a U-Net design, and the model is trained by minimizing the L2 distance between an original image and a restored one at the pixel level:

$$\mathcal{L}_{SSL} = ||X - \hat{X}||_2, \tag{1}$$

where X is an original input image and $\hat{X}$ is a restored image by the model.

4. Implementation Details

4.1. Pre-training protocol

Our SSL model is trained exclusively on *unlabeled* CT cardiac volumes. To ensure no test-case leaks from proxy tasks to target tasks, any volumes that will be used for testing in target tasks are excluded from pre-training. Following (Haghighi et al., 2021), 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. The SSL loss function is based on mean

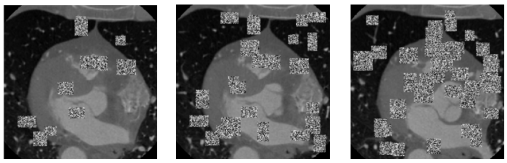

Figure 3: An illustration of the masking module's output that is utilized in our SSL framework. Moving from left to right, the task difficulty progressively increases, resulting in greater distortion of the input. The leftmost image corresponds to level 1, the middle image to level 2, and the rightmost image to level 3.

squared error (MSE). The input volumes are resized to 256×256×16, and we apply min-max normalization to standardize the 3D imaging volumes. This normalization ensures that pixel values are within an HU range between -1024 and 2000, effectively clipping any values that fall below or above this range. We modify the last layer of the decoder to reconstruct the image rather than a segmentation mask in the proxy task. We consider different masking techniques as shown in Figure 2. For the level 1, 2, and 3 of curriculum learning, we mask out 25, 30, and 35 blocks, respectively, with a probability of 0.8. We utilize the minimum 8×8 pixels, and the maximum 16×16 pixels for block's spatial sizes. The masking block sizes and locations are randomly selected. We use AdamW optimizer (also considered RMSprop, the further detail is available in the next section) 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.

4.2. Fine-tuning protocol

We utilize a 3D U-Net network for cardiac structures segmentation task. During this phase, the encoder is initialized with the pre-trained encoder through our SSL approach, while the decoder is randomly initialized due to the substantial differences between the target segmentation and the proxy reconstruction tasks. Furthermore, all the downstream model's parameters are fine-tuned. We use AdamW optimizer (also considered RMSprop) with a learning rate

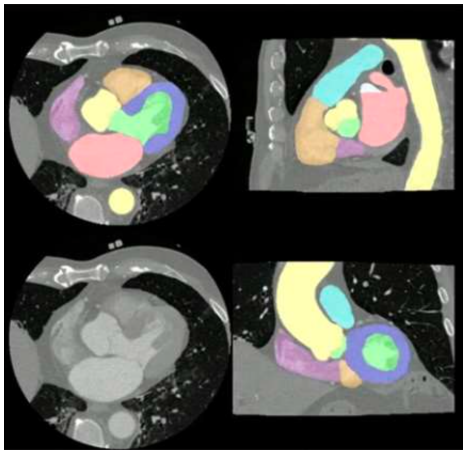

Figure 4: An example of a CT cardiac chamber image annotated by clinical experts. Each unique color represents different heart substructures, including the left atrium (LA), left ventricle (LV), right atrium (RA), right ventricle (RV), myocardium (MYO), aorta (AO), pulmonary artery (PA), and left atrial appendage (LAA).

of 0.001. To prevent over-fitting, we employ early-stopping technique with a patience of 50 using 10% of the training data as the validation set. We evaluate the segmentation performance using the Dice coefficient. We adopt training from random initialization, which is the prevailing approach in 3D cardiac CT segmentation methods, as the standard baseline for comparison. For fair comparisons, the baseline and our method benefit from the same data and settings in the training and testing phases.

4.3. Data preparation

In this study, we use an in-house dataset of 3D CT cardiac imaging. The dataset has been collected from 32 different hospitals in 10 different countries worldwide, introducing inherent challenges such as heterogeneity and distribution shifts stemming from the utilization of different imaging scanners. Additionally, the dataset exhibits a broad diversity in patient cohorts and is characterized by class imbalance. As such, the dataset under the study represents a significant level of diversity, serving as a valid indicator for evaluating the generalizability of our pretrained model. The dataset includes a total number of 262 3D scans obtained from 262 patients. The dataset includes a total number of 65,418 slices. The dataset provides pixel-wise segmentation masks for eight heart substructures, including 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 all 262 cardiac CTA series (see Figure 4 for an example of annotation results by clinical experts). The size of each 3D image volume is 512×512 with varying numbers of images in the z-direction, ranging from 140 to 560 with a median of 224. Each image volume is normalized to $[0, 1]$ by -1000 Hounsfield Unit (HU) and +2000 HU. We randomly divide the dataset into train, validation, and test sets, including 168, 43, and 51 volumes, respectively. It should be noted that the test set is an independent set that is not used in pretraining and fine-tuning stages. The information about the cardiac substructures data distribution is available in the appendix.

4.4. Data sampling strategy

The scarcity of training data, coupled with the pronounced class imbalance in the cardiac CT datasets, hinders the training of deep models with broad generalizability for cardiac segmentation task. To overcome this challenge, we propose a data sampling strategy to augment the quantity and diversity of training and validation data. Figure 5 illustrates an example of the proposed sampling strategy. Typically, the initial and final slices within a 3D image volume tend to carry less crucial information, whereas the middle slices frequently encompass the most information for diagnosis tasks. This pattern is often observed in medical imaging, particularly in CT cardiac scans, where human hearts are often situated in the middle slices within each image volume, whereas the initial and final slices often carry less relevance to cardiac issues, containing limited information for model training. Consequently, it holds greater promise to selectively sample the more informative slices, primarily situated in the middle of the CT volume. To achieve this, we first divide the slices of each 3D volume into three distinct buckets. Subsequently, we randomly sample 60 3D image patches, each sized 512×512×16, from every CT volume in training and validation sets. Notably, for each CT volume, slices from the middle bucket (bucket 2) are sampled twice as frequently as slices from the other two buckets. Table 1 summarizes the number of

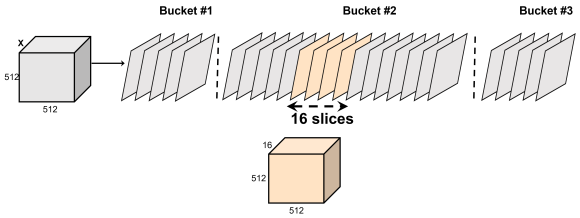

Figure 5: Data Sampling Strategy for Expanding Training Data and Mitigating Class Imbalance. Images from each 3D volume, which may have varying z-dimensions, are categorized into three buckets. This strategy enables the selection of 512×512×16 3D image sub-volumes, with an emphasis on the second bucket. This approach not only addresses class imbalance but also ensures compatibility with GPU memory.

training and validation data after applying our proposed sampling strategy. These samples are used during the pretraining and fine-tuning stages.

Table 1: Data set details after the sampling process.

Data split	Sub-volumes	Ratio
Train	10,080	80%
Validation	2,580	20%

5. Experimental results

The following delves into the cornerstone of our experimental results.

5.1. Cropping strategy: Random vs. Center Crop

Given the dynamic nature of human heart structures, both proxy and target tasks can be sensitive to the choice of hyperparameters during training. Consequently, it is crucial to explore various hyperparameter configurations to achieve optimal performance for cardiac CT segmentation task. As such, we conduct experiments to examine the impact of multiple cropping strategies as well as crop sizes on the target task performance.

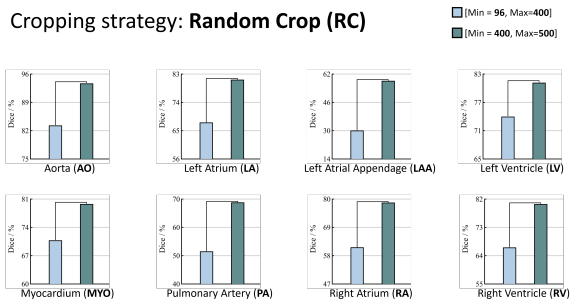

Figure 6: Impact of Random Crop with Varying Sizes on Segmentation Performance of Each Cardiac Structure: Larger cropping sizes significantly outperform smaller cropping sizes for each cardiac structure.

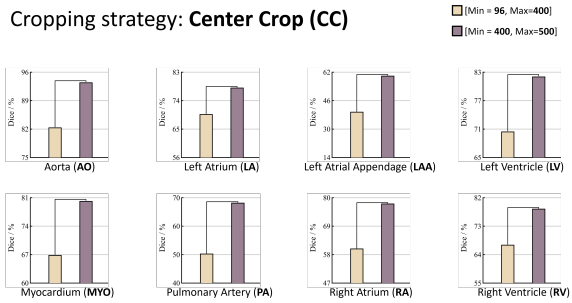

Figure 7: Impact of Center Crop with Varying Sizes on Segmentation Performance of Each Cardiac Structure: Larger cropping sizes significantly outperform smaller cropping sizes for each cardiac structure.

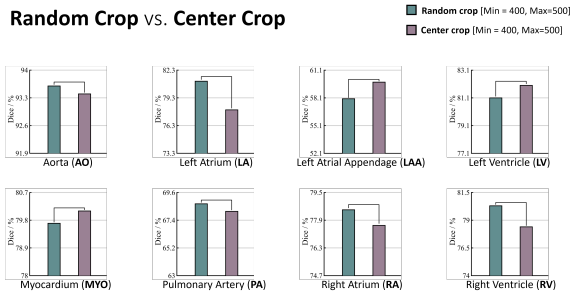

Figure 8: Impact of Cropping Strategy: Random Crop outperforms Center Crop in the segmentation of the majority of heart substructures.

First, we delve into the impact of cropping size on the performance of CT cardiac segmentation tasks. We examine two cropping strategies, including Random Crop (RC) and Center Crop (CC). We specifically explore two distinct scenarios for both random and center cropping strategies. In the first scenario, we span a range of crop sizes, from small to large, where the crop sizes are randomly selected from within the interval of [96, 400]. In the second scenario, we exclusively opt for larger crop sizes, randomly chosen from the range of [400, 500]. In both scenarios, we employ a uniform distribution to randomly generate crop sizes for both center and random crop strategies. Figure 6 illustrates the impact of crop size when employing the random cropping strategy. Notably, the crop size emerges as a pivotal factor influencing target task performance, with consistently lower performance observed for crop sizes in the [96, 400] range compared to larger crop sizes in range [400, 500]. Figure 7 depicts the impact of crop size when employing the center cropping strategy, confirming a similar observation: larger crops in the [400, 500] range consistently deliver superior performance across all chamber structures.

We further investigate the impact of cropping strategy, specifically comparing random and center cropping, on the performance of CT cardiac segmentation tasks. In this experiment, we employ the best-performing crop size range (i.e., crops with a minimum size of 400 and a maximum size of 500) for both random and center cropping. As observed in Figure 8, in the majority of cases, random cropping surpasses the performance of center cropping. The superior performance of random cropping can be attributed to its ability to generate a more diverse set of training samples during the training process. Additionally, since the cropping operation varies in each epoch due to the use of random numbers, a greater variety of cases is introduced during the training phase, contributing to the observed improvement in performance.

5.2. Optimizer: RMSprop vs. AdamW

We investigate the impact of different optimizers on the segmentation performance of each cardiac structure. To do so, we examine two optimizers, including RMSprop (Ruder, 2016) and AdamW (Loshchilov and Hutter, 2019). As seen in Figure 9, AdamW provides superior performance over RMSprop, highlighting its potential to be considered as a proper optimizer for the cardiac CT segmentation tasks.

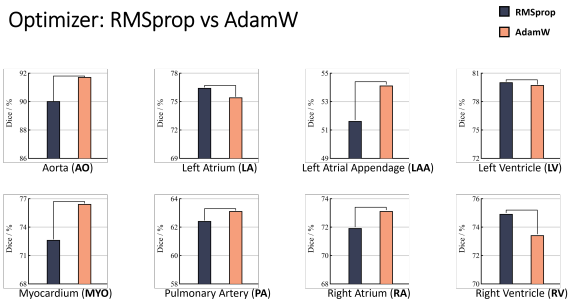

Figure 9: Impact of Optimizers: AdamW outperforms RMSprop in the segmentation of most heart substructures, underscoring its potential as a promising optimizer for CT cardiac image segmentation of heart substructures.

5.3. Impact of Self-supervised Learning

To demonstrate the effectiveness of our developed 3D curriculum SSL method, we adopt training from random initialization, which is the prevailing approach in 3D cardiac CT segmentation methods, as the standard baseline for comparison. Figure 10 illustrates the results of fine-tuning our pretrained model for cardiac CT segmentation versus training the target model from random initialization.As seen, our developed 3D curriculum SSL method consistently outperforms the baseline across all cardiac structures. Particularly, in challenging cardiac structures, such as LA, LAA, RA, PA, and RV, the gap between our method and the baseline becomes more pronounced. Our results demonstrate the effectiveness of our method in capturing generalizable representations capable of effectively segmenting various cardiac chamber structures.

5.4. Ablation studies

Impact of Task Difficulty Level. We examine the influence of varying pretraining task difficulty levels on target task performance. In particular, starting from level 1, we progress through level 4 while employing two masking strategies: Random Noise and Zero Masked. Figure 11 displays the target task performance at different pretraining difficulty levels for

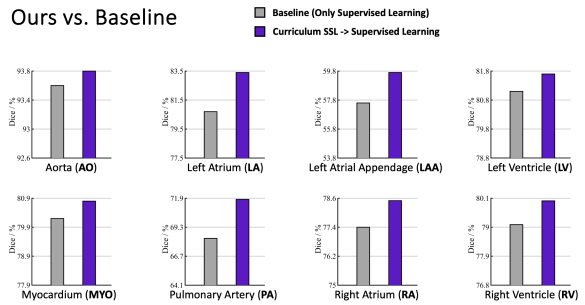

Figure 10: Our proposed self-supervised method consistently outperforms the baseline in the segmentation of all eight heart substructures.

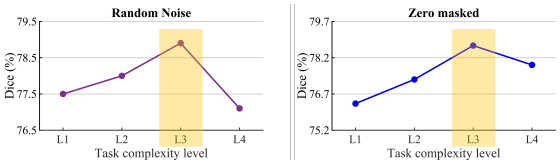

Figure 11: Ablation Study on Task Complexity Levels during the Pre-training Stage: As task complexity increases, the performance improves for both Random Noise and Zero Masked strategies, highlighting the effectiveness of our learning strategy. Notably, the optimal complexity level in both scenarios occurs at L3, suggesting that excessively challenging tasks may negatively impact representation learning.

both strategies. Notably, in both Random Noise and Zero Masked scenarios, performance begins to decrease after Level 3. This could be attributed to the increased difficulty at Level 4, where substantial image distortion impedes the model's ability to capture meaningful representations, resulting in degraded target performance.

Impact of Different Masking Strategies. We investigate the effect of different masking strategies in our self-supervised design. To do so, we examine the three different masking techniques as presented in Figure 2. Figure 12 illustrates the target task performance for different masking strategies in our self-supervised learning framework. We draw the following observations from the results. First, the mean value masking module provides the lowest

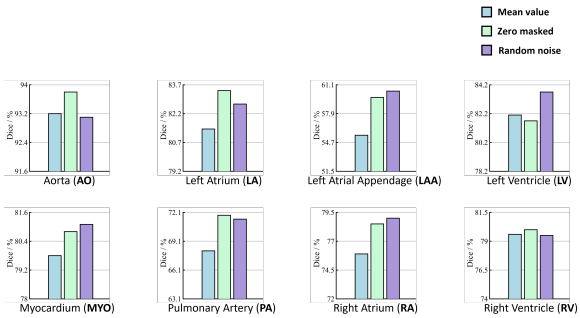

Figure 12: Ablation Study on Different Masking Strategies: The Mean Value strategy doesn't significantly enhance our SSL model's ability to capture more generalizable representations. In contrast, both the Zero Masked and Random Noise strategies exhibit competitive performance compared to each other.

performance among all three different masking techniques. Intuitively, it appears that when using the mean masking strategy, the model may learn shortcut solutions, rather than capturing informative features for solving the reconstruction task at hand. This can be attributed to the local image continuity, where nearby pixels often share similar content and have similar pixel intensities, approximating their mean value. Consequently, the model may take a shortcut by predicting this mean value which minimizes the objective function, but potentially neglecting important image details and nuanced patterns throughout the images. This oversight significantly hinders the model's ability to extract meaningful representations effectively, thus impacting its overall performance. Moreover, the Random Noise and Zero Masked alternatively provide the best performance in different chamber structures.

6. Conclusion and Future Work

In this paper, we propose a new Self-Supervised Learning (SSL) method for CT cardiac image segmentation. Our proposed SSL method is designed based on the curriculum learning paradigm and masked image modeling. In particular, we gradually increase the levels of complexity in the masked image modeling proxy task. Initially, lightweight distortions are applied to images, allowing the model

to learn general knowledge from images and serve as effective contextual clues for solving more complicated tasks during subsequent pretraining stages, resulting in more generalizable features for cardiac CT imaging. We also propose and examine different masking strategies for self-supervised pretraining, including Mean masking, zero masking, and random noise masking. Our experimental results show that our curriculum SSL technique improves the performance of the CT cardiac image segmentation task in each cardiac chamber. Additionally, our ablation studies suggest that as task complexity increases in SSL, the performance improves, highlighting the effectiveness of our learning strategy. Moreover, our results suggesting that excessively challenging tasks may negatively impact representation learning. Although our SSL method was evaluated for CT cardiac imaging, it is generic and applicable to semantic image segmentation tasks in various medical imaging modalities. As part of future research, we aim to extend our SSL approach to other imaging modalities, including MR and Ultrasound imaging.

Acknowledgments

We are grateful to Gopal Avinash, Sandeep Dutta, Amy Deubig and Yunfeng Li at GE Healthcare for their valuable comments and suggestions to make this work possible.

References

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. 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.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7, 2020a. ISSN 2297-055X. doi: 10.3389/fcvm.2020.00025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti 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.

James D. Dormer, Ling Ma, Martin Halicek, Carolyn M. Reilly, Eduard Schreibmann, and Baowei Fei. Heart chamber segmentation from CT using convolutional neural networks. In Barjor Gimi and Andrzej 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.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal 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.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming 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.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B. Gotway, and Jianming 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.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20824–20834, June 2022.

Fatemeh Haghighi, soumitra ghosh, Sarah Chu, Hai Ngu, Mohsen Hejrati, Han Hui Lin, Baris Bingol, and Somaye Hashemifar. Self-supervised learning for segmentation and quantification of dopamine neurons in parkinson's disease. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Joseph Harms, Yang Lei, Sibo Tian, Neal S. McCall, Kristin A. Higgins, Jeffrey D. Bradley, Walter J. Curran, Tian Liu, and Xiaofeng 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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross 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.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B. Gotway, and Jianming 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.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B. Gotway, and Jianming 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.

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

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017, sep 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Shadab Momin, Yang Lei, Neal S McCall, Jiahan Zhang, Justin Roper, Joseph Harms, Sibo Tian, Michael S Lloyd, Tian Liu, Jeffrey D Bradley, Kristin Higgins, and Xiaofeng 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.

Eric D. Morris, Ahmed I. Ghanem, Ming Dong, Milan V. Pantelic, Eleanor M. Walker, and Carri 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.

Partha Pratim 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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro 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.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali 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.

Qianqian Tong, Munan Ning, Weixin Si, Xiangyun Liao, and Jing 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.

Chunliang Wang and Örjan 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.

Jing Wang, Shuyu Wang, Wei Liang, Nan Zhang, and Yan 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, 2022a. doi: https://doi.org/10.1002/acm2.13597.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022b. doi: 10.1109/TPAMI.2021.3069908.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han 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.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming 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.

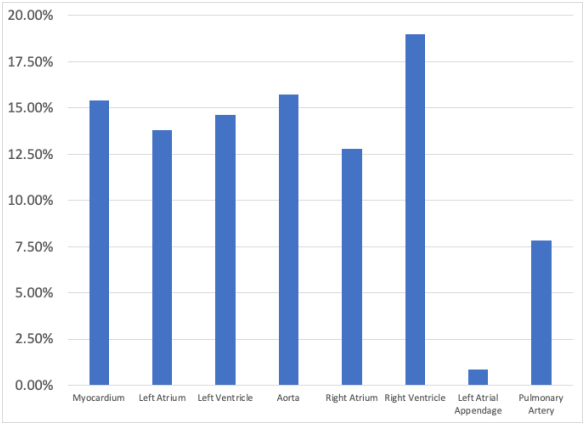

Figure 13: The cardiac substructure data distribution in the data set.

Appendix A. Appendix

Figure 13 shows the cardiac substructure data distribution in the data set we use in this paper. Each cardiac substructure has roughly equal ratios of pixel values except the Left Atrial Appendage (LAA). As we discuss in the introduction section, the LAA is the most difficult substructure to segment by a computational method because of the size of the anatomy.