Eötvös Loránd University

Faculty of Informatics

Dept. of Artificial Intelligence.

Myocardial Perfusion Imaging using Vision Transformers

Supervisor:
Szűcs, Ádám István
PhD Candidate

Author:
Haris Ali
Computer Science MSc

Budapest, 2025

Declaration

I hereby declare that this thesis titled "Myocardial Perfusion Imaging with Vision Transformers" and the work presented in it is my own original research. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at Eötvös Loránd University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given.
- I have acknowledged all main sources of help.
- This thesis has not been submitted for any other degree or professional qualification.

Signed: Haris Ali

Date: 28th April, 2025

Contents

Abstract

The manual delineating the left ventricle (LV) in Myocardial Perfusion Imaging (MPI) is one of the most labor-intensive and time consuming tasks in nuclear cardiology and radiology. The outcome of the diagnosis of the MPI is extremely dependent on the accuracy and the consistency of the segmentation of the ventricles, hence the process is done under extreme caution in order to minimize the risks of any possible error. However, the process of turning this task into an automated one present a number of challenges that need to be mitigated. Fist of all, the Signal-to-noise ratio (SNR) is mostly low and the resolution of the image is limited, complicating the process of detecting the boundaries. Secondly, the high disparity in both the cardiac traces uptake ad the differences in the hardware used for the imaging introduces inconsistencies. Finally, their is a lack of a standardized definition of the shape of the LV and the there is no standard shape that can be traced based purely on image data, which introduces a lot more ambiguity in the task.

This thesis proposes a novel method built to address the limitations mentioned above by using a Transformer-based architecture, integrating Statistical Shape Prior (SSP) technique. This approach is specifically used to mitigate the data-hungry nature of the transformers in case of limited data. The proposed architecture achieves over 4% improvement over a number of metrics in segmentation and classification against the bench-marked State-of-the-Art (SOTA) approaches use for LV segmentation, both on the synthetic data and the real-world clinical scans.

In addition to the improvements in the quantitative metrics, the incorporation of the prior shape information enabled the model to learn insights into the variability and the structural patterns of the LV anatomy in MPI Single-Photon Emission Computed Tomography (SPECT) imaging. This deeper understanding of the LV enhances the reliability of the AI-powered automatic segmentation of the LV and also the general comprehension of the morphology of the LV in clinical practice.

Chapter 1

Introduction

MPI using SPECT pllays an important role in the process of non-invasive assessment of the Coronary Artery Disease (CAD). Considering cardiovascular diseases being one of the leading causes of mortality all across the world, the need for an efficient, accurate and accessible tool for diagnosis is at a high demand. MPI SPECT provides critical information about the perfusion status of the heart, which helps in the early detection and planning the treatment which improves the outcomes of the patients.

Radionuclide MPI under a specific condition, such as stress, is majorly regarded as one of the most effective diagnosis technique, which is also non-invasive, in order to identify the or detect the CAD. Using the application of MPI SPECT, clinicians become equipped to diagnose and detect the functionally relevant coronary stenoses with a relatively high level of specificity This actually enables them to make decisions that are informed and possibly the right ones regarding the pathways of the patients' treatment [1]. By visualizing the perfusion process of the heart muscles, clinicians can detect the areas of the heart where there is a presence of coronary stenoses or obstructions which may be the causing issue for inducible perfusion deficits under the conditions of stress or rest. This ability of diagnosis is not only essential to identify the patients with CAD but also functions as an important tool for mitigating patient risk and guiding the decision making process of the clinicians.

MPI using SPECT has emerged as both an effective and economically viable modality for the purpose of diagnosis. MPI based SPECT offers both the advantages of being accessible and having established standard clinical protocols hence it is the preferred choice of a number of diagnostic processes. One of the major strengths of

MPI is the adaptability of the technique, as it can incorporate a number of radiopharmaceutical agents, such as 201Tl Chloride, 99mTc Tetrofosmin, and 99mTc Sestamibi, which basically is dependent upon the imaging protocols and imaging needs. The mentioned agents are typically administered intravenously before the image acquisition part, and then the collected image data are later reconstructed using techniques which are dedicatedly designed for cardiac imagery. The last, and possibly the most crucial, stages in the diagnostic process involves the segmentation of the anatomical structures relevant to the diseases and then the reorientation of this segmented volumetric data. This part of the diagnostic is usually performed by trained clinical professionals in order to ensure precision, better reliability and to mitigate the risks of errors.

Beyond the usage of the perfusion imaging alone, there are additional functional parameters, which are valuable, that can be derived when gated acquisition techniques are applied. These parameters include End-Systolic Volume (ESV), End-Diastolic Volume (EDV), and the left ventricular ejection fraction (LVEF). All of the mentioned parameters are essential in order to indicate the performance of the heart. The values of these parameters are computed through the precise delineation of the myocardial boundaries of the LV,, which makes the task of segmentation even more crucial in the whole pipeline. The perfusion and the functional analysis collectively provide a detailed understanding of not only the vascular but also the mechanical health of the heart.

Efficient and accurate quantitative analysis of the 3D MPI SPECT data is extremely sensitive to a number of factors that are involved in the full end-to-end imaging and reconstruction pipeline, as mentioned above. All of these factors together contribute not only to the reliability of the evaluation of the data, but also to the detection of a range of cardiac abnormalities [2]. The important step in this process is the segmentation and reorientation of the LV, which basically refers to the determination of the spatial alignment of the LV and its segmentation based on the anatomical midline. The tasks of both reorientation and the segmentation within MPI SPECT imaging have been acknowledged, for a long time, as one of the central yet difficult challenges. Over the course of years, multiple commercial systems have been developed in order to counter these issues, but more often relying on very extensive and curated datasets in order to ensure reliable performance in clinical environments [3], [4], [5].However, the existing solutions fall short when they are

applied to the newer reconstruction paradigms, especially in the situations where there are only a limited number of labled patients datasets. In order to mitigate these imitations faced by the current solutions and to increase the generalization capabilities of the models under limited data conditions, approaches incorporating self-supervised learning and few-shot learning have gained popularity. Nevertheless, the effectiveness of these strategies is most of the times overshadowed by the high costs associated with the expert annotations. In addition to this the lack of consensus regarding a standard segmentation protocol also complicate the practical application of the processes.

In the recent years, within the field of MPI SPECT imaging, the adoption of Deep Learning (DL) techniques are looking at a significant revival [6]. This renewal is basically driven in part by the development of the novel radio-tracers and also the growing clinical demand to minimize the amount of administered radiation dose and also the image acquisition time of the performed procedures [7]. As a consequence, the modern methods of reconstruction have been focusing on configurations that are based on low photon count data, sparse acquisition views and reduced amount of injected doses [8], [9], [10]. But in-spite all that, the advancements do not fully resolve the challenges which are inherent to the segmentation tasks of MPI SPECT. Despite using SOTA neural network based reconstruction strategies, the segmentation accuracy is still heavily relied on the underlying reconstructed images. When working with lower-dose inputs, the images mostly lack proper structural clarity, which diminishes the benefits which are offered by the DL based reconstruction methods. Even in situations where the image reconstruction achieves are visual equivalence to a full-dose filtered back projection methods, the issues of low SNR, Poisson noise characteristics, and the impact of partial volume effect (PVE) continue to affect the generalization capabilities and hence the reliability of automated segmentation models.

In this work, is proposed a novel approach in order to solve the aforementioned bottlenecks of the segmentation task, all the while also contributing further detailed insights into the anatomical characteristics of the MPI SPECT LV. Contrary to the previous approaches employed for the task, where the use of isolated pre-processed regions, or usage of cropped volumes, is common, the method in this study makes use of the entire reconstructed image volumes, hence incorporating all of the contextual spatial cues which are available within the full field-of-view (FOV). The choice of this

design makes sure that no information that is diagnostically relevant is discarded, hence allowing the model to infer the left ventricular structure in relations to the surrounding regions of the anatomy. This holistic approach increases the robustness of the model, specifically in cases where abnormalities in the patterns could possibly interfere with the more localized analysis.

In order to mitigate the limitations hat are associated with Convolutional Neural Network (CNN), specifically their receptive field being restricted which hinders them from learning long-range dependencies, the proposed method employs a fully transformer based architecture called nnFormer [11]. This architecture is specifically developed for tasks pertaining to volumetric medical imaging. It allows the network to learn global reasoning over the 3D structures which offers a significant advantage over the traditional CNNs in situations where the boundaries of the organs are not sharply defined such as SPECT. But there is a limitation to using transformer architectures, which is that they require a huge amount of data in order to learn acceptable global representations and have a good generalization ability. Hence, in order to overcome such a limitation, the proposed method incorporates SSP as a regularization technique. Such shape priors introduce an anatomical consistency into the DL model which acts as a guidance signal during the training of the model. This approach helps the models in situations where the amount of available data is limited. Using the shape priors, the model is made to learn meaningful and spatially coherent segmentation outputs even with minimal amount of supervision. This whole process bridges the gap between the traditional rule-based segmentation models and the fully data driven DL approaches.

Chapter 2

Related Work

The automatic reorientation and segmentation process of MPI SPECT represent steps that are essential for the accurate and and efficient diagnosis and the quantitative analysis of the heart. A number of commercial softwares have been developed in order to perform these tasks and are widely adopted in clinical practices. The Corridor4DM [12], which was developed at the university of Michigan ,provides a platform for the comprehensive quantitative analysis for Myocardial Perfusion and the functional assessment from SPECT. This extremely integrated system gives access to an automated processing tool for analysis and reporting which si specifically developed to meet the increasing demands of such tools. In a similar way, the Emory Cardiac Toolbox (ECTb) [13] implements an extensive pipeline of quantitative tools which are developed as a result of very extensive research and its validation. It features a database of normal perfusion with more than 150 patients each, the Fourier analysis of regional thickening, used for functional assessment, and a very advanced display function that allows to display 3D volumes for image fusion. The Cedars-Sinai approach [14] focuses on an end-to-end automatic expert system which is based on mathematical algorithms and rules based on logic reasoning. The presented QGS software is been used at more than 20 thousand locations all across the globe. The yale method [15] is focused on the quantification process of both the MPI and specifically the LV functional abnormalities, which address the challenges faced by the process by multiple factors such as background of images and the defects of perfusion using specialized processing techniques.

Other than the mentioned commercial solutions for the tasks, there is significant amount of research efforts that have been devotedly carried out to develop more ad-

vanced approaches and algorithms for MPI SPECT segmentation and reorientation. The Level-Set Methods (LSMs) have been proved to me one of the most developed methods in the field. The research by [16] presented an automatic method for the segmentation of the LV based on variational level sets in volumetric SPECT. This method integrates adaptive thresholding for the estimation of initial closed curves, followed by the evolution of variational level set for the determination of the final contours. Very effective performance has been demonstrated using this approach as compared to the manual delineation through ROC analysis. More advanced LSM techniques [17] developed model for implicit level sets representations which is based on 4D statistical shape analysis that combined the temporal information gotten from gated SPECT sequences. This eliminated the need for challenging point correspondences, at the same time outperforming 3D models with a better characterization of the evolution of the temporal shape.

Multiple hybrid approaches have also been developed in order to address some specific challenges in MPI SPECT analysis. The charged contour model presented in [18] is designed specifically to handle the concavities in the segmentation label volumes. Later, the research conducted by [19] proposed a novel approach that combined the shape and appearance priors using a constraint with levels set deformable models, hence implementing a soft-to-hard probabilistic constraint that provided a lot more flexibility as compared to rigid shape constraints alone. This approach proved to be particularly effective for LV segmentation in 4 dimensional gated SPECT, even if there were perfusion defects present. The study presented in [20] developed hybrid active contour model for Myocardial D-SPECT volumes thar combined local image fitting models with the region-scalable fitting energy functions in order to mitigate the inhomogeneity issues, all the while maintaining the computational efficiency. More earlier work by [21] developed a statistical model-based approach with the usage of 3D Active Shape Models (ASM) which combined both the geometric shape and the information depicted by the gray-level appearance from training data for the purpose of achieving robust segmentation of gated SPECT MPI.

Despite all the advances in the research, the traditional approaches still continue to face great challenges in order to achieve the globally optimal point especially when processing the complete FOV volumes with a varying amount of image quality and variability in the anatomy. Such limitations has driven the more recent explo-

ration of the machine and DL techniques in the field of nuclear cardiology. Early machine learning applications in the domain were specifically focused on sub-tasks of the whole segmentation pipeline. The research done by [22] showed the effectiveness of using support vector machines (SVMs) for predicting the optimal valve positioning, demonstrating that the approach can be comparable to expert performance in SPECT alignment, at the same time reducing the dependence on users for quantification. This study highlighted how even the conventional machine learning approaches can improve some specific aspects of the workflow of cardiac analysis. The study done in [23] presents a comprehensive review showing that DL solutions have shown remarkable promise across multiple aspects of PET and SPECT imaging, from the quantitative analysis to the instrumentation part of it all. Another study in [24] presents a specific discussion about the transformative impact of using CNNs on the task of LV segmentation, showing their ability to learn and interpret complex features directly from images.

The dawn of DL in the world brought forward even more comprehensive solutions to the task of LV segmentation. Another research [25] proposes an end-to-end fully CNN based architecture that directly learns the segmentation mapping from the SPECT images taken as input. This approach eliminated the need to process the volumes in multiple stages for the segmentation map. This way the approaches using DL in order to handle the entire segmentation task could be developed, replacing the traditional way, with a unified framework. Building on the same idea [26] implemented a U-Net [27] based CNN architecture that outperformed significantly their own previous dynamic programming solution presented in [28], and it particularly improves handling the complex variations of shape of the LV myocardium. More recently, there have been an increasing number of studies in even more sophisticated network architectures that are tailored to some of the unique challenges that are prevalent in the analysis of cardiac SPECT. Moreover, research in [29] introduced convolutional long-short term memory (Long-Short Term Memory (LSTM)) units in the skip connections of a V-Net architecture [30], which enabled an effective way of extracting temporal features from gated SPECT sequences. This novel approach addressed an essential need to leverage the presence of temporal information in higher dimensional volumes of SPECT. In a similar way, [31] enhances the usual 3D U-Net architecture with a self-attention mechanism which is employed at the bottleneck. This allows for better inclusion of the global contextual information throughout the

volumetric data.

The usage of shape priors have been identified as one of the efficient strategies, valuable for improving the accuracy of segmentation models. The study by [32] proposes a method that includes the shape priors, which are generated using a dynamic programming algorithm into a 3D V-Net network using a spatial transformer network (STN) which proved to give extremely good results al the while maintaining anatomical consistency. Despite the research, there is still a lot of unexplored room when it comes to shape priors and their use in the segmentation process [25]. The existing solutions rely heavily on a number of factors. The first one being the availability of a huge dataset size available with labels in order to train a model. Secondly, the focus within the architecture have been CNN and thirdly, even the use of hybrid mechanisms which include attention systems have at least one convolutional layer component. While being effective, these approaches most of the time require substantial training data or very complex model architecture that might limit clinical applicability.

As opposed to the previous works in the field, the method proposed in this study introduces a number of innovative ideas. Firstly, the DL model used in the study is a fully transformer based architecture, which diverts from the traditional approaches of using convolutional blocks. Using this approach allows to capture long-range global dependencies in the input 3D volumes. Moreover, and more importantly, our approach incorporates the statistical shape priors in a way that compliments the strength of the transformer model for better performance. This approach achieves performance that is comparable to the SOTA methods, all the while requiring half the amount of training, with limited data, addressing a probllem that is critical in the clinical deployment where there is not a huge amount of annotated dataset available. The novelty of this research is built upon, with substantial extension of the previously present work. Even though [33] demonstrates the value of using spatial transformers and and [32] discuss te usage of shape priors, this method unifies the approaches within a single transformer framework. Compared to the attention method presented in [31], the approach presented in this study is more comprehensive self-attention based paradigm throughout the whole network. The gains in the efficiency, relative to [34] are very noteworthy, as a strong performance is achieved her without the need of a self-supervised pretraining phase.

Chapter 3

Methodology

3.1 Overview

The main goal of the research is to propose a comprehensive and strong method which is designed specifically to significantly improve the accuracy of the segmentation of MPI using SPECT for the LV. The precise segmentation of MPI SPECT images is extremely critical for the detection and the assessment of CAD. However, the task of achieving a high accuracy in segmentation poses a number of challenges due to a multitude of inherent limitations of MPI SPECT data, such as low SNR partial volume affects, substantial noise because of the Poisson statistics, motion artifacts and, obviously, the anatomical variability among different patients.

In order to address these challenges, the proposed research amalgamates advanced approaches in the field of DL, more specifically utilizing the transformer based architecture known as nnFormer [35], which is combined with an innovative idea of using SSPs. The nnFormer architecture is chosen because of the ability of it of capturing both the local and the global information or contextual relationships in volumetric data as opposed to traditional CNNs which only capture local information. nnFormer leverages the Local Volume-Based Multi-head Self-Attention (LVMSA) and the Global Volume-Based Multi-head Self-Attention (GVMSA) mechanisms very efficiently in a unified method. These modules of the transformer architecture very effectively encode the long-range dependencies which are extremely important for better segmentation accuracy specifically in medical imaging where they are characterized by indistinct boundaries.

Simultaneously, the SSP are merged into the segmentation pipeline in order to

improve the anatomical consistency of the model. SSPs provide the DL model with a mathematical model which captures the probabilistic variability off the LV that are derived from the data annotated by experts. This SSP methods employs an advanced technique of optimization such as Mahalanobis distance based regularization and the Kullback-Liebler (KL) divergence in order to refine the segmentation boundaries. This way the outputs of the segmentation model maintain plausible anatomical outputs, which improves the segmentation accuracy even when the input data is incomplete or ambiguous. The combination of nnFormer and SSP provides us with a novel hybrid architecture. The full architecture pipeline is presented, on an abstract level, in fig. 3.1.

This hybrid approach leverages not only the strengths of DL models in extracting complex and hierarchical feature representations from volumetric data, but also the advantages of SSPs in maintaining consistent anatomies. This approach also addressees the limitations of the existing methods, which include inefficient generalization capability and dependencies of large, precisely annotated data for training. It also mitigates the impact of a number of different imaging artifacts and the noise, which enhances the overall reliability on the segmentation.

Extensive procedures for training involving efficient optimization strategies such as Adam, and specifically segmentation tailored loss such as the DiceCELoss, are implemented in order to ensure stable performance across a diverse set of patients. The training and the validation aspects are conducted rigorously using an extensive dataset of MPI SPECT which consists of diverse collimation methods and demographics of the patients which improves the generalization capability of the method. Extensive setups of computation which leverage high performance GPU computing environments ensure efficient training and inference of the model. Comprehensive evaluation metrics are utilized in order to quantitatively validate the performance of the segmentation. These metrics include precision, recall, intersection over union (IoU) and Dice coefficient. These metrics provide an extremely in-depth insights into the capability of the method to handle real-world variability and complex scenarios.

In a summary, the proposed methodology contributes to not only a significant advancement in the division of cardiac image segmentation but also provides a practical and robust solution which is applicable in a clinical environment. The combination of nnFormer and SSP ensures reliable and precise segmentation having clinically meaningful results, which paves the way for better and improved diagnosis and patient

outcomes in CAD management.

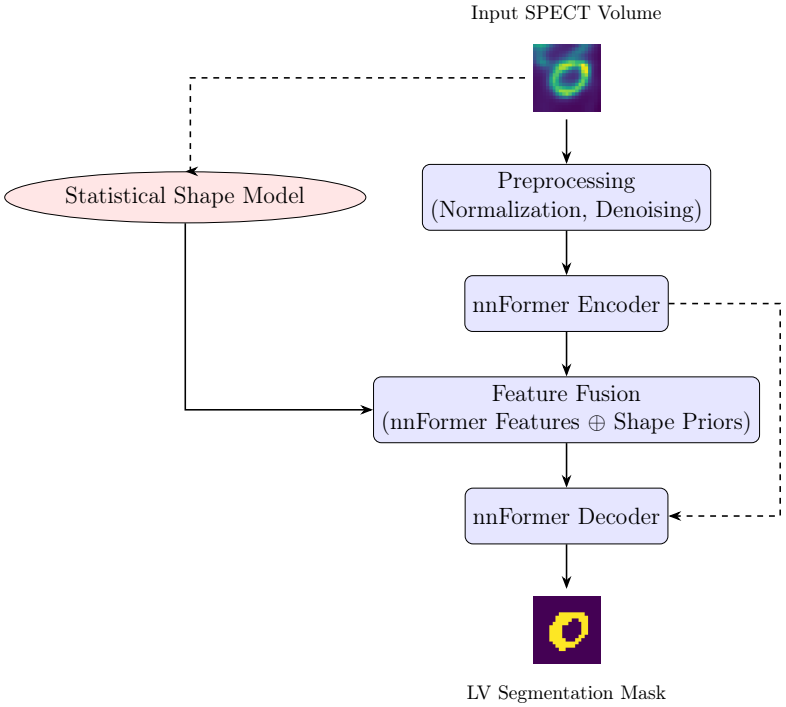

Figure 3.1: Network architecture pipeline

3.2 Data Acquisition and Preprocessing

The MPI dataset which is utilized in this research was acquired using SPECT. This acquired dataset consists of volumes from a total of 80 patients, which are carefully selected in order to represent the diverse demographic and the characteristics of the clinic. The population of the patients included individuals with varying age groups, physiological conditions and gender distributions in order to ensure the reliability, robustness and generalizability of the model. Multiple different radio-tracers were employed in the process of acquiring the MPI SPECT, specifically agents labled by technetium-99m(Tc) such as Tc Tetrofosmin and TC Sestamibi, and also the thallium 201 chloride (T1 Chloride). Each of these radio-traces offer unique properties in imaging thereby providing a comprehensive coverage of all the possible clinical scenarios which are encountered in everyday diagnostic.

The dataset of the patients used in the research is divided into five distinct groups. **(A)** The first, and the largest group is a heterogeneous black-box dataset which comprises of patients who are scanned under multiple d efferent geometries, pharmaceutical protocols and the settings of the acquisition. This group consisted

of 40 individuals, out of which 27 were female and 13 male with an average age of 69.62 years. The average height of the group is 167.25cm while the average weight is 78.11 kgs. **(B)** The second group consists of 10 patients who are imaged using the most recent MPH collimator configurations,, specifically the APT73 collimators installed on the Mediso AnyScan Trio system. This group consists of 7 females and 3 males with an average age of 68.8 years, an average height of 169 cm, and an average weight of 73.8 kgs. **(C)** The third group comprises of another 10 patients scanned with the LEHR-HS collimator which is also mounted with a trio camera. Out of these, 8 are female and 2 male with an average age of 68 years, an average height of 176 cm and an average weight of 92 kgs. **(D)** the fourth group was scanned using the earlier generation imaging hardware, in order to provide a comprehensive basis with the legacy systems. This group included data from 10 patients who are scanned using the Mediso CardioD system in a seated configuration setting. This group consisted of 1 female and 9 male subjects with an average age of 75.6 years, an average height of 151 cm and a mean weight of 71 kgs. **(E)** The fifth and last group also involved 10 individuals who are imaged with the Mediso CardioC system with subjects positioned supine. This group consists of 7 females and 3 males. The retrospective nature of this dataset is stored in the interfile format due to format limitations and only the average age could be reliably extracted which comes out ot be 70.33 years.

Each of the patient went through very rigorous imaging procedures which are adhering strictly to standard protocols of acquisition in clinics. The patients were administered the mentioned radio-pharmaceuticals intravenously which was followed by image acquisition after the standardized waiting period that allows sufficient tracer uptake in the myocardial tissue. The image acquisition protocols were varying based on the collimation method which was employed. For example imaging with the MPH collimator a very specific step-and-shoot helical trajectories, on the other hand the stationary collimator positions were employed for the other collimators which creates different spatial sampling patterns and different challenges to image reconstruction. After the acquisition of the raw data, a number of precprocessing techniques were employed in order to prepare the data for the subsequent segmentation analysis. The preprocessing pipeline was developed in order to address a number of inherent issues with the imaging and to optimize the data quality for better segmentation outcomes.

The preprocessing steps began with the correction of the attenuation utilizing the TeraTomo reconstruction algorithm [36], which majorly removed the attenuation artifacts which are caused by the soft bone and tissue structures. This step is very essential in order to ensure the uniformity in the distribution representation of the tracer across the myocardial tissue, hence improving the segmentation accuracy. In some cases where the attenuation correction data was not available, an Ordered Subset Expectation Maximization (OSEM) algorithm [37] was used in order to reconstruct the full FOV volumes, providing us with data with robust handling of Poisson noise and preserving essential details of the image.

In order to improve the image quality even more, noise reduction techniques are employed, specifically targeting the reduction of the Poisson noise which is the most prominent in MPI SPECT imaging due to the low count of the photons. More advanced filtering methods were also employed such as the Gaussian smoothing and the adaptive median filtering in order to balance the noise reduction with the preservation of important boundaries anatomically and the structural details. The partial volume effects (PVE), which majorly has an impact on the accuracy of the segmentation because of blurring tissue boundaries, were addressed systematically using dedicated PV correction techniques and de-convolution techniques. These methods restored the sharpness in the images and enhanced the delineation of th myocardial boundaries, especially in the regions which have complex anatomical structures.

The images that are the result of the above preprocessing are made to go through further normalization procedures in order to ensure consistency in the scales of intensity across all the datasets which helps in having more robust training of the final segmentation models. Standardization of the intensities of the voxels involved scaling the pixel intensity distribution in order to have a mean of zero and a unit variance, which significantly improves the numerical stability, which in-turn helps the convergence of the DL models.

All the data preprocessing steps are performed in a very structured and repeatable framework, using scripts that are custom developed in Python and specialized libraries for numerical computation, imaging and DL such as PyTorch [38], NumPy [39] and Scikit-learn [40]. The comprehensive documentation of the preprocessing parameters and the configurations was nicely maintained so as to ensure the transparency and the reproducibility of the methodology. The final dataset after the preprocessing provided us with a very high-quality and standardized input data for

the training, validation and the testing of the DL models. The careful handling of the whole data acquisition pipeline ensuring the variability and the rigorous preprocessing ensured optimal MPI SPECT image preparation which majorly enhanced the accuracy and the reliability of the segmentation which was obtained from the hybrid model.

3.3 Detailed Description of nnFormer Architecture

The nnFormer architecture introduces an innovative advancements in the field of medical image segmentation, which is specifically designed in order to address the limitations that are faces by the traditional CNNs. nnFormer does it by efficiently capturing both the local and the global spatial relationships in the volumetric medical data. This section gives a very detailed description of the architecture explaining the components, their integration, and the rationale behind the usage of the components in the specified manner.

3.3.1 Overall Architecture

The nnFormer architecture follows a structure that is very much used in the image segmentation field. It follows a U-shaped encoder decoder architecture which is inspired by the widely used U-Net. This choice of the structure helps in efficient learning of the detailed local features, all the while preserving and leveraging the global contextual information across multiple scales of resolution. The nnFormer architecture comprises of three main components: an encoder, a bottleneck and a decoder which are all interconnected via skip connections.

3.3.2 Encoder

The encoder of nnFormer, as shown in fig. 3.2 starts with an embedding layer, which consists of multiple convolutional layers with very small kernel sizes, typically using 3x3x3. This is followed by Gaussian Error Linear Units (GELU) activation and then ends with a normalization layer. This initial convolutional based embedding transforms the input volume into a higher dimensional feature space which encodes the low-level spatial detailed, which are necessary for subsequent processing, very efficiently. After the embedding layer, the encoder uses a LVMSA blocks. These

LVMSA blocks are designed so that they can capture the local spatial dependencies within the segmented volumes, which significantly reduces the computational complexity of the model as compared to the conventional global self-attention mechanisms.

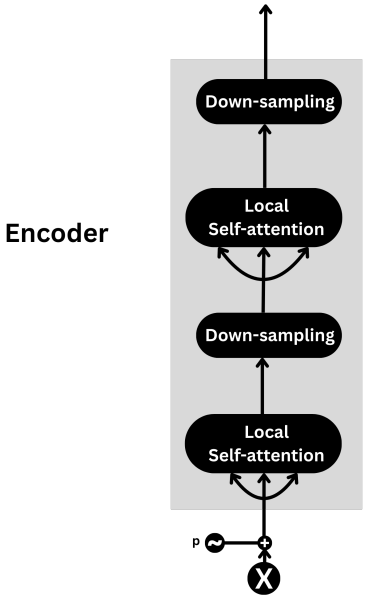

Figure 3.2: Encoder of nnFormer

Each of the LVMSA blocks is further comprised of successive layers of transformer modules in order to use attention mechanisms to effectively model the very intricate local interactions contextually. The encoder also combines very strategically placed downsampling convolutional layers which reduces the spatial dimensions of the feature maps while at the same time progressively increasing the depth of the feature map. This downsampling process helps the extraction of the hierarchical features in a more abstract way and on a global scale based representations at lower resolutions, which are essential for capturing the broader variations and anatomical structures.

3.3.3 Bottleneck

In the center of the nnFormer there is a bottleneck ,shown in fig. 3.3. This bottleneck has GVMSA (GVMSA) mechanisms. Contrary to the LVMSA, the GVMSA provides with a significantly bigger receptive field which captures the long-range dependencies across the whole global context of the volumetric feature map. This increased area of the receptive field is essential at this stage in order to allow the

network to achieve a comprehensive understanding of the global representation and the anatomical structures, improving the overall segmentation accuracy. The bottleneck very effectively combines the complex spatial dependencies and also the high level features which are extracted by the encoder. This serves as a robust foundation for the decoder for accurate and consistent output during the decoding.

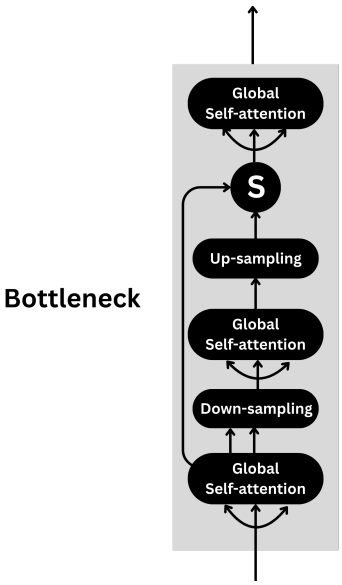

Figure 3.3: Bottleneck of nnFormer

3.3.4 Decoder

The decoder, in fig. 3.4 is, in a way, mirrored version of the encoder it also employs LV-MSA blocks but coupled with convolutional upsampling, instead of the downsampling, or so-called transposed convolution. This restores the spatial resolution of the feature maps gradually to the original dimensions of the input. Each step of the upsampling process in the decoder is designed so as to reconstruct the detailed anatomical information by combining the high resolution detail captured spatially from the corresponding encoding stages via skip connections.

A prime innovation of the nnFormer is the use of the skip attention mechanisms, in place of he traditional concatenation or summation which is typically used in skip connection mechanisms. These skip connections very selectively integrates the features of the encoder with the corresponding features of the decoder, which are guided by the attention weights that highlight relevant spatial features dynamically

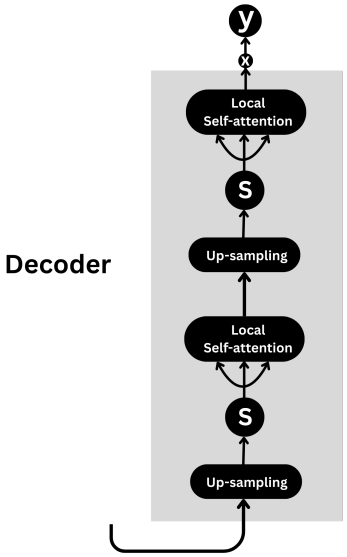

Figure 3.4: Decoder of nnFormer

and hence suppress the irrelevant features. This selective combination majorly improves the precision of the segmentation, specifically in the areas where the clear anatomical delineation is very challenging due to the noise and artifacts.

3.3.5 Attention Mechanisms

nnFormer utilizes two different types of attention mechanisms: the LVMSA and GVMSA. LVMSA very efficiently models the local spatial dependencies by partitioning the feature maps into manageable patches of volumes, which reduces the computational complexity without sacrificing a lot of the performance of the attention mechanism. In contrast to this, GVMSA models the global interactions spatially across the entire volumetric feature maps, which are essential for capturing the large scale anatomical structural integrity and the contextual relationships. Both of the attention mechanisms use multi-head configurations which enable parallel computations of attention across multiple representational subspaces. This multi-ha design greatly improves the capability of the network in order to concurrently capture very diverse spatial relationships and the interactions at multiple scales hence thereby improving the segmentation accuracy.

3.3.6 Integration and Optimization

The amalgamation of LVMSA, GVMSA and the convolutional operations in the nnFormer is very carefully optimized in order to leverage the strength of each of these methods. The convolutional layers provide the efficient encoding of the low level spatual features, while the LVMSA and the GVMSA collectively capture both the complex spatial contexts and the long-range dependencies which are crucial for the precise and robust segmentation, providing us with the overall nnFormer architecture shown in fig. 3.5. Optimization of the nnFormer involves a specialized loss function called the Dice cross entropy loss (DiceCELoss). This loss function objectivises a high segmentation accuracy while maintaining a reliable realistic plausibility, which effectively guides the learning process towards more generalizable models across diverse imaging conditions.

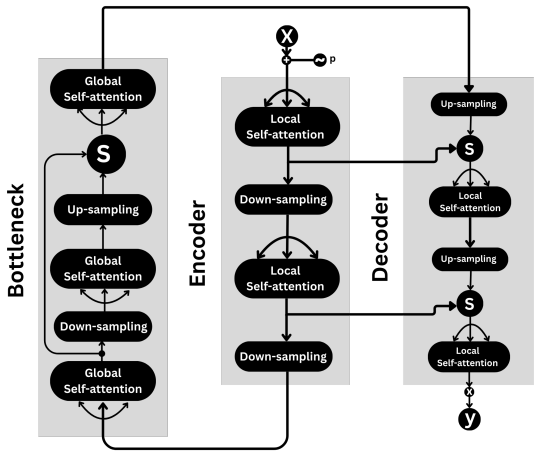

Figure 3.5: Full nnFormer architecture

In conclusion, the nnFormer architecture represents a very sophisticated segmentation framework which is specifically tailored for medical imaging systems. The innovative design of the architecture and the advanced methods of attention overcomes the traditional limitations of segmentation models by effectively using optimization techniques across components ensuring accurate, robust and clinically meaningful segmentation outputs.

3.4 Statistical Shape Prior

SSP offer a very powerful approach in order to enhance the robustness of the segmentation, specifically within the domain of medical imaging having data that might have sparse annotations available and noise embedded into it. In the context of MPI SPECT, SSPs are employed as a form of regularization technique that combined the prior knowledge about the left ventricular shape or its anatomy directly into the learning process of a DL model. This section provides a comprehensive examination of the SSP method which is used in this research including its theoretical, background, implementation and integration within the DL pipeline.

3.4.1 Motivation for SSPs

Tradition DL approaches often rely completely on the pixel-level intensities and their patterns in order to learn the segmentation boundaries. However, using this strategy can become very unreliable in low quality image settings, where the tissue boundaries are most of the time blurred or occluded due to the low resolution of images, noise in them, and partial volume effects. In order to address this issue, SSPs use a population level anatomical information in order to put a constraint on the segmentation process for plausible output shapes increasing the resilience to noise and variability. In MPI SPECT, the consistence delineation of the LV is extremely critical for precise and accurate computation of the functional parameters of the cardiac imagery The shapes of the LV vary across different patients but it maintains a stable topology that can be modeled statistically. SSPs capture this anatomical prior and provide a probabilistic framework that biases the training process towards valid segmentation outputs.

3.4.2 Shape Prior Construction

In order to contruct the shape priors, a mathematical model for the LV is used which is used to generate a featr space. The input SPECT volumes are first aligned to a common coordinate system using techniques such as affine or non-rigid registration. After the alignment, the shape representations are extracted, commonly as binary masks. In this work, the shapes are encoded using contoured vectors in a latent space that is suitable for probabilistic modeling. Once the shapes are aligned,

Principal Component Analysis (PCA) is applied to the shape representations in order to identify the primary modes of variation. This gives out a low-dimensional shape space where each shape can actually be represented as a linear combination of the mean shape and a set of principal components. The statistical model can be expressed as:

$$\mathrm{S} = \bar{S} + \mathrm{Pb}, \tag{3.1}$$

where $\bar{S}$ is the mean shape, P is the matrix of principal components, and b is a vector of shape parameters.

The parameters b follow a Gaussian distribution which is estimated from the volumes, which forms the basis of the shape prior and is used later to asses the possibility of a predicted shape.

3.4.3 Regularization Using Mahalanobis Distance

In order to integrate the SSP into the learning, a shape regularization term is added to the loss. This term penalize the deviations from the learned shape space using Mahalanobis distance, which is defined as:

$$D_M(b) = (b - \mu)^T \Sigma^{-1} (b - \mu), \tag{3.2}$$

where μ and Σ are the mean and covariance of the shape parameters from the training data.

The constraint forces the network to produce output shapes that lie in the distribution of the known shapes. It acts as a force that corrects and guides the model to produce good LV geometries.

3.4.4 KL Divergence-Based Optimization

In addition to the Mahalanobis distance, a Kullback-Lieber (KL) divergence term is used in order to further align the predicted distribution of shapes with the prior one. The KL divergence quantifies the difference that exists between the predicted shape distribution $a(b)$ and the prior distribution $p(b)$:

$$D_{KL}(q||p) = \int q(b) \log \left(\frac{q(b)}{p(b)} \right) db. \tag{3.3}$$

This term is particularly useful when training any probabilistic model such as a variational autoencoder (VAE) in order to learn to generate a whole distribution of anatomical shapes.

3.4.5 Integration into the Segmentation Network

The integration of the SSP into the segmentation network follows a specific architectural strategy, instead of applying the regularization on the standalone loss term, the key novelty lies in embedding the priors directly into the feature space of the network, which influences the decoder using enhanced intermediate representations.

The process begins with sampling the shape prior from the learned statistical distribution which is conditioned on the features that are derived from the input volume of SPECT. This priors reflects a anatomically likely structure that is aligned with the patient's scan, and selected from a low-dimensional latent shape space. For each of the sampled prior there are two metrics that are calculated The first is the Mahalanobis distance (D_M) and the second is the KL divergence (D_{KL}), which quantify how well the prior conforms with the expected shape. These metrics are then combined into a single scalar shape prior loss:

$$\mathcal{L}shape = \lambda_1 D_M + \lambda_2 DKL, \tag{3.4}$$

where λ_1 and λ_2 are empirically tuned coefficients.

Instead of using this $\mathcal{L}_{shape}$ as a regularization term in the loss, first the derivative of this loss is calculated which gives a tensor providing the derivative on the surface of shape prior. This derivative is then reshaped into a tensor matching the spatial dimensions of the nnFormer bottleneck output. Then this reshaped loss derivative term is concatenated to with the bottleneck of the feature map. This results in a fused representation that encodes both the contextual features that are data driven and the shape information giving the anatomical constraints. The decoder then processes this combined tensor, enabling the final segmentation to benefit from the shape priors. This integration guides the feature propagation throughout the whole network, improving the spatial coherence and the consistency of the segmentation.

3.5 Training and Optimization Procedures

The training and optimization of the full segmentation framework were specifically designed to ensure the convergence and the generalization capability of the model across varying anatomies of the patients and different imaging conditions. The architecture was implemented using the PyTorch library and trained on high-performance Nvidia GPUs, leveraging both the nnFormer and the SSP modules.

3.5.1 Training Dataset and Splits

The training dataset consisted on MPI SPECT scanes from 60 different patients These scans were carefully selected in order to ensure the variability in the type of the tracer used the quality of the image and the features of the anatomy. The validation/testing dataset consisted of 14 different scans from different patients. The stratification ensured proper and proportional representation of the collimator types and the demographics of the patients across split.

3.5.2 Data Augmentation

In order to improve the generalization capability and avoid over-fitting to the training data, several different data augmentation techniques were employed during the training phase. These include:

- Random rotation
- Spatial padding (To ensure consistent input size)
- Random cropping

All of the augmentation or transformations were applied in real time using the Monai library in order to maintain anatomical plausibility and preserving the label integrity.

3.5.3 Loss Function

The loss function that is used in the study is the DiceCELoss known as the Dice cross-entropy loss, which combines the functioning of both the dice coefficient and the cross entropy loss. This formulation focuses on 2 crucial aspects of the

medical image segmentation, which are class imbalance and probabilistic boundary confidence.

The dice component of the loss function directly optimizes the function for the overlap the predicted labels and the ground truth which makes it very effective for the datasets that target the structures that occupy a small portion of the volume, which is what exists in myocardial SPECT. Meanwhile the cross-entropy term ensures that the voxel-wise confidence of the classification is properly incorporated, allowing the whole network to learn the fine-grained details and maintain the sharp boundaries needed.

This hybrid loss is proved to be both differentiable and computaitonally efficient, with facilitates stable convergence during the training process. This loss was selected over te traditional single term losses due to the fact that its robustness in handling small, complex anatomical data even in the presence of noise. By relying completely on the the DiceCELoss, the training pipeline remains completely streamlined and very effective, which eliminates the need for additional tuning of the hyperparameters that is associated with the multi-loss formulations.

$$\text{DiceCELoss} = \left(1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i}\right) - \sum_i g_i \log(p_i) \tag{3.5}$$

In the above formulation, p_i represents the predicted probability for voxel i, while g_i denotes the corresponding ground truth label, which is typically binary. The index i runs over all voxels (or pixels) in the image or volume. In the case of multi-class segmentation, the formulation extends to include an additional class index c, where $p_{i,c}$ indicates the predicted probability that voxel i belongs to class c, and $g_{i,c}$ is the one-hot encoded ground truth label for voxel i with respect to class c. The terms $\sum_i p_i$ and $\sum_i g_i$ represent the total predicted and ground truth positive volumes, respectively, while $\sum_i p_i g_i$ measures the overlap between prediction and ground truth. The cross-entropy component, $\sum_i g_i \log(p_i)$, quantifies the voxel-wise classification error by penalizing deviations between predicted probabilities and ground truth labels. Together, the Dice coefficient term and the Cross-Entropy loss term jointly capture both region-level overlap and voxel-level classification confidence, promoting accurate and stable segmentation outputs.

3.5.4 Optimization Strategy

Model training used the Adam optimizer [41] with the following configuration:

- Initial learning rate: 1×10^{-5}
- Learning rate decay: exponential decay with a factor of 0.01 every 10 epochs
- Batch size: 2 volumes per iteration
- Weight decay: 1×10^{-6}
- Number of epochs: 100 for nnFormer baseline, 50 for proposed SSP-integrated model

The best performing model, based on the validation loss, is saved for final testing.

3.5.5 Evaluation Pipeline

Following the training, model performance was evaluated on the test set. For each volume of the patient, the output of the segmentation is generated in a single forward pass The predictions were post-processed using the morphological operations in order to remove isolated false positives and maintain region continuity. Quantitative metrics such as Dice coefficient, recall, precision, and F1 score were computed.

3.5.6 Implementation Environment

All of the training and evaluation procedures were conducted in a Linux based environment with CUDA-enabled GPUs. the full code was implemented in Python using PyTorch, with additional support of libraries such as monai, pandas, numpy and matplotlib fro visualisation. The rigorous training and the optimization pipeline ensured the resulting model was both the accurate and very generalizable which is robust against the variability of imaging, and is efficient enough for potential deployment in clinical settings.

3.6 Implementation Details and Computational Environment

In order to ensure the reproducibility,, optimal training efficiency and the scalability the whole segmentation framework was implemented using an extremely module DL environment. This section details the stack of the software, the computational infrastructure and the engineering strategies which are adopted in order to support the development and the training, validation and testing phase.

3.6.1 Software Framework

The model is developed using Python, leveraging the DL library PyTorch 1.12 version, due to its flexibility and the wide adoption of the library in the research community all across the computer science community. Other key supporting libraries used in the research are:

- **Monai** for data loading, augmentation, and patch-based processing.
- **NumPy** for numerical operations.
- **Scikit-learn** for evaluation metrics and other machine learning tasks.
- **Matplotlib and Seaborn** for visualization of training curves and result analysis.
- **Pydicom and Nrrd** for medical image I/O, including support for DICOM and nrrd formats.

3.6.2 Hardware Infrastructure

The training of the model was performed on a server which is equipped with NVIDIA GTX 1080Ti GPU. These resources enabled efficient handling of very large scale volumetric datasets and helped with Simultaneous experimentation. The training sessions were accelerated using:

- CUDA 12.8 for GPU-accelerated matrix operations.
- cuDNN to optimize the neural network routines.

- PyTorch's automatic mixed precision (AMP) in order to reduce the usage of memories but without sacrificing the accuracy of the model.

3.7 Evaluation Metrics and Experimental Setup

In order to correctly evaluate the performance of the proposed methodology for segmentation ,a very comprehensive set of quantitative metrics is utilized, together with a carefully structured experimental setup. These selected metrics were chosen in order to capture both the geometric consistency of the segmentation and also the anatomical plausibility across different imaging conditions.

3.7.1 Evaluation Metrics

The evaluation of the segmentation performance focused on the following key performance metrics:

- **Dice Similarity Coefficient (DSC):** Measures the overlap between the predicted and ground truth segmentations, defined as:

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|}, \tag{3.6}$$

  where P and G denote the predicted and ground truth segmentations, respectively.

- **Intersection over Union (IoU):** Provides an alternative measure of overlap, calculated as:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}, \tag{3.7}$$

  which balances sensitivity and specificity by considering both false positives and false negatives.

- **Precision:** Indicates the proportion of predicted positive voxels that are truly positive, defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.8}$$

  where TP is the number of true positives and FP is the number of false positives. High precision implies fewer false positive segmentations.

- **Recall (Sensitivity):** Measures the ability of the model to correctly identify all relevant voxels belonging to the target structure, given by:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3.9}$$

where FN is the number of false negatives. A high recall value indicates effective capture of the entire structure, even if some false positives occur.

These metrics were computed for all volumes in the test set and averaged to provide mean performance indicators, standard deviations, and confidence intervals.

3.7.2 Experimental Protocols

The multiple experiment-based configurations were employed in order to analyze the robustness and the generalizability of the proposed architecture:

1. **Baseline Comparison:** The base nnFormer architecture was trained and evaluated independently to establish a performance baseline.

2. **Shape Prior Augmentation:** The proposed integrated model combining nnFormer with SSP was evaluated to quantify the impact of incorporating anatomical priors into the segmentation pipeline.

3. **Comparison with SwinUNETR:** The SwinUNETR model, using the same preprocessing and hyperparameter configurations as nnFormer, was trained and evaluated to provide a comparative benchmark.

4. **Noise Robustness Analysis Using Shape Priors:** Robustness to image degradation was tested by introducing Poisson noise to phantom images generated from the shape prior model. These phantoms served as synthetic input SPECT volumes, allowing evaluation of the model's ability to handle severely noisy conditions.

5. **Low Data Regime Simulation:** Subsampling experiments were conducted to evaluate segmentation performance under limited labeled data conditions (using 10, 20, 30 40 and 50 patients) and the performances for both the nnFormer and the nnFormer+SSP were compared.

These experimental evaluations comprehensively demonstrate the efficiency, the clinical reliability and the generalization capability of the nnFormer integrated with the SSP model across a ranging MPI SPECT scenarios.

3.8 Summary and Justification of Methodology

The choices made about the methodology in this study were guided by the combined objectives of achieving SOTA accuracy of segmentation and also ensuring the robustness of the method under constraints in a clinical setting, which include limited annotated data and high variance in the MPI SPECT image quality. Each component of the proposed architecture was selected specifically in order to address the challenges that are observed in the previous studies done in medical image segmentation approaches. The adoption of nnFormer as the backbone of the whole architecture is justified because of its ability to model global spatial dependencies and also the contextual relationships, which are most of the times very crucial for resolving ambiguities in a low resolution an high noise volumetric scans. Unlike the conventional CNN-based architectures operating usually with limited receptive fields, structure of the nnFormer based on transformers allows the model to learn long-range anatomical correlations within the entire volume.

Despite the advantages offered by the transformer-based architectures such as the nnFormer they still require a large amount of data to learn effectively. In order to overcome this limitation and enhance the generalizability considering low data settings, the SSP integration was introduced in the architecture. This update helps the model b embedding domain knowledge into the DL model training phase as a way of providing anatomical regularization. This method provides a constraint to the segmentation model's output to plausible shapes, hence enhancing the reliability especially in cases that are challenging to work with suc as scans having perfusion defects or motion artifacts. The inclusion of this SSP focuses on an important gap that exists in the existing methods. Traditional CNNs, and even transformer models can produce segmentation that might be structurally not plausibile, especially when faced with data that is noisy or sparse. By incorporating the shape regularization, which is based on the Mahalanobis distance and the KL divergence penalty, the proposed methodology ensures that the model adheres to expected anatomical con-

figurations, but not at the cost of flexibility in the learning of the data. This balance is very critical in order to maintain credibility in real-world clinical applications.

In addition to all of this, the loss function, so-called the DiceCELoss, was motivated by the need of balancing the pixel-wise accuracy with the anatomical correctness of the output. This hybrid loss contributes to having consistency during the convergence process during the training phase, all the while promoting accuracy of the segmentation in both common and edge-case scenarios. Another very important decision was the use of multiple experimental protocols including the baseline comparisions of nnFormer and the SwinUNETR with our proposed architecture, and noise robust testing. All of these experiments were structured to validate the performance improvements and also to assess the generalizability of the model across multiple different imaging conditions. From the perspective of implementation, the usage of high-performance GPU hardware and a scalable software design ensured that the model is very efficiently trained and tested.

Chapter 4

Results

The evaluation of the segmentation methodology that is proposed in this research was structured in a very progressive and the systemic manner which begins with the evaluation using synthetic phantom data and then extending into a very comprehensive analysis on real patient datasets. This layered evaluation approach allowed for a very detailed inspection of the model's behavior under controlled and also some clinically realistic conditions, which validates its robustness, generalization capability and the precision. The results were assessed using a number of statistical and validation techniques.

4.1 Evaluation on Synthetic Phantom Data

The first phase of the evaluation process involves the testing of the segmentation pipeline on simulated data, or so-called the phantoms, by introducing Poisson noise into the data using th x-cat phantom generator [42]. Such a type of phantom is very widely regarded for the anatomical realism and is very frequently used in nuclear medicine as a gold standard baseline for the validation of the method. In this experiment varying levels of the Poisson noise were synthetically added to proper clean phantoms in order to simulate differing SNR conditions. The rationale behind the usage of Poisson noise is rooted in the fact that the nature of the SPECT imaging physics, where noise originates rom stochastic processes during the photon detection. This proves Poisson noise to be an appropriate and clinically relevant choice for the performance benchmarking of the models. The noise levels spanned a range from the severely degraded (low SNR) to relatively clean (high SNR), providing deep

insights into the robustness of the algorithm against various levels of deteriorated image quality.

The quantitative analysis of the model on phantoms revealed that the proposed model maintains a performance above a random baseline across all the tested conditions but more importantly, it shows a notable increase in the segmentation accuracy, particularly after the 0 dB SNR as showed in fig. 4.1. While all the metrics showed an improvement with increasing SNR, one very significant result was the behavior of the precision of the model across the range. The precision showed 3x to 4x improvement compared to the other available metrics, showing that the model provides strong in selectivity in identifying relevant regions even in noisy volumes. This capability is very essential for clinical reliability, where the false positives could lead to unnecessary procedures.

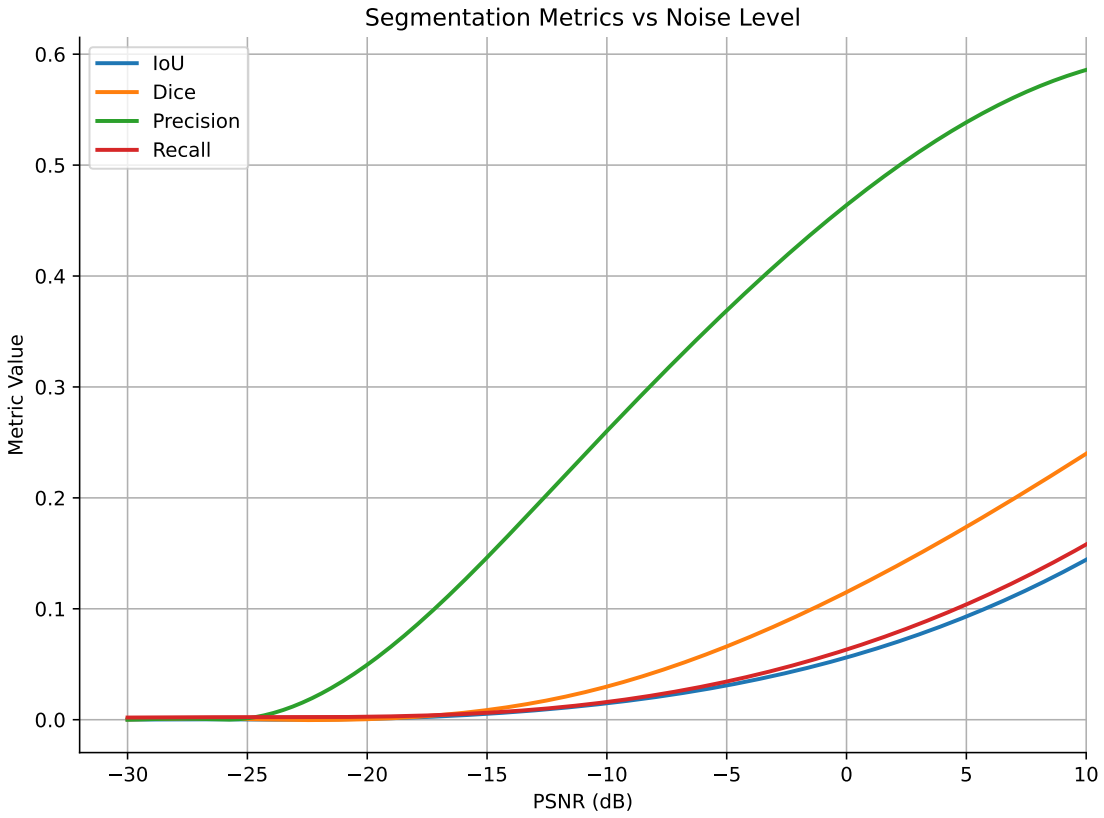

Figure 4.1: Metric values against Noise in phantoms

4.2 Quantitative Comparison with Transformer Architectures

In order to benchmark the proposed architecture against the SOTA transformer based segmentation models we implemented two relevant and efficiently proved alternatives: nnFormer [11] and Swin-UNETR [43]. All of the models were trained under identical conditions using the same 60 patient training size and the 14 patient validation size in order to ensure a fair comparison between the models. The averaged quantitative results are summarized in Table 4.1. The proposed architecture consistently outperformed both the models for comparison across almost all the evaluation metrics. notably the precision, dice and the IoU score were higher which reflects an enhanced segmentation accuracy and spatial consistency. The one exception was the recall, where Swin-UNetR achieves a slightly higher values due to the consistent over-segmentation trend that is observed over several different samples. This suggests that while Swin-UNetR may be able to capture more positive instances, it does that at the cost of specificity which leads to a higher false positives rate. These results prove that the superior balance that needs to be achieved between the precision and the recall is done by the architecture proposed, which is an extremely important aspect of the clinical segmentation tasks in order to avoid both under and oversegmentation.

Averaged performance metrics				
	Precision	Recall	IoU	Dice score
Our model	**0.714**	0.7545	**0.5706**	**0.7172**
nnFormer	0.6819	0.6457	0.4715	0.6334
SWIN-UNetR	0.5329	**0.9158**	0.4949	0.6404

Table 4.1: Averaged segmentation results on the patient dataset. The proposed shape prior enhanced transformer is able to outperform the nnFormer [11] and SWIN-transformer [43] approaches in most metrics.

4.3 Effectiveness of Shape Priors on Anatomical Conformance

More in-depth investigation into the role of SSP was conducted through a feature space analysis which is visualized in fig. 4.2. Contrary to the unbiased transformer models that rel completely on the features hierarchies that are learned, this method benefits from anatomical guidance which is embedded completely through the shape prior network. The visualizations confirm that the incorporation of prior-based optimization majorly improves the ability of the model to predict the boundaries of the LV. The optimization with this constraint which is introduced by the shape prior encourages better outcomes anatomically and also provides a better regularization for the learning process. These findings confirm the hypothesis that the use of SSP can guide the structural learning of the model resulting in better performance and easier interpretibility.

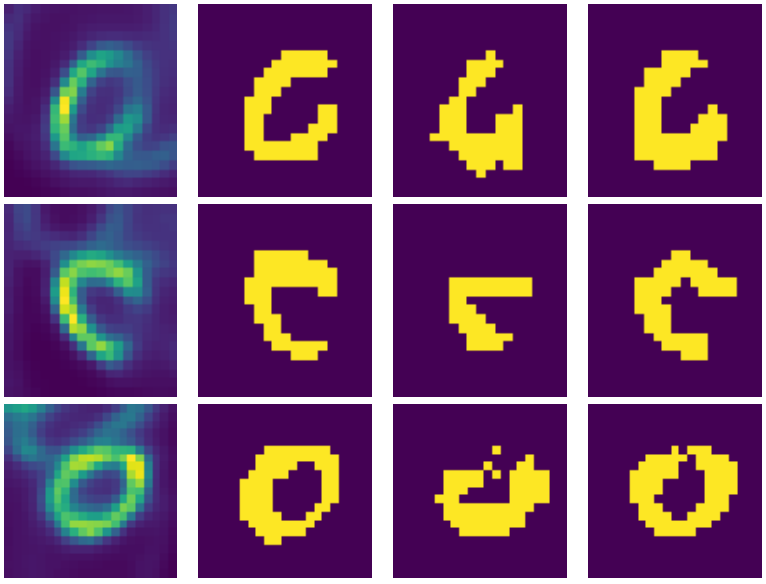

Figure 4.2: Segmentation results, the first column depicts the source LV volume, the second column is the ground truth labels. The third column shows the best performant competing model the nnFormer and the last is the proposed model's predicted LV labels

4.4 Receiver Operating Characteristic (ROC) Analysis

In order to assess the behavior of the segmentation model as a classifier-like model, a Receiver Operating Characteristic (ROC) curve was calculated on the test set of the patients, which is shown in fig. 4.3. ROC is a very well established method which helps to evaluate the discriminative power of a model. It works by plotting the true positive rate against the false positive rate across a number of thresholds. An interesting finding is the fact that the curve reveals a small fluctuation in the performance, which a one point dips below the threshold of the random classifier, which is 0.5, after the 0.75 mark. Upon very close examination of the model, this drop is found to be connected to the inconsistencies and the imperfections in the ground truth labels which were manually annotated and hence got subject to inter-observer variability.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{4.1}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{4.2}$$

This observation highlights the very essential role of the having the phase of quality control in the preparation of the dataset. Inconsistent labeling of the ground truths can lead to misleading evaluations, which strengthens the need for a standardized protocol of labeling or the usage ot multi-annotator consensus labels The ROC curve does not just evaluate the behavior of the model but also surfaced the underlying limitations that are present in the dataset, which adds value beyondd the traditional diagnostic use.

4.5 UMAP Embedding of Bottleneck Representations

In order to gain further insights into the learned representations of the developed model, we performed a method of dimensionality reduction called the Uniform Manifold Approximation and Projection (UMAP) [44] on the bottleneck features that are extracted from the bottleneck as an output. We select five samples at ran-

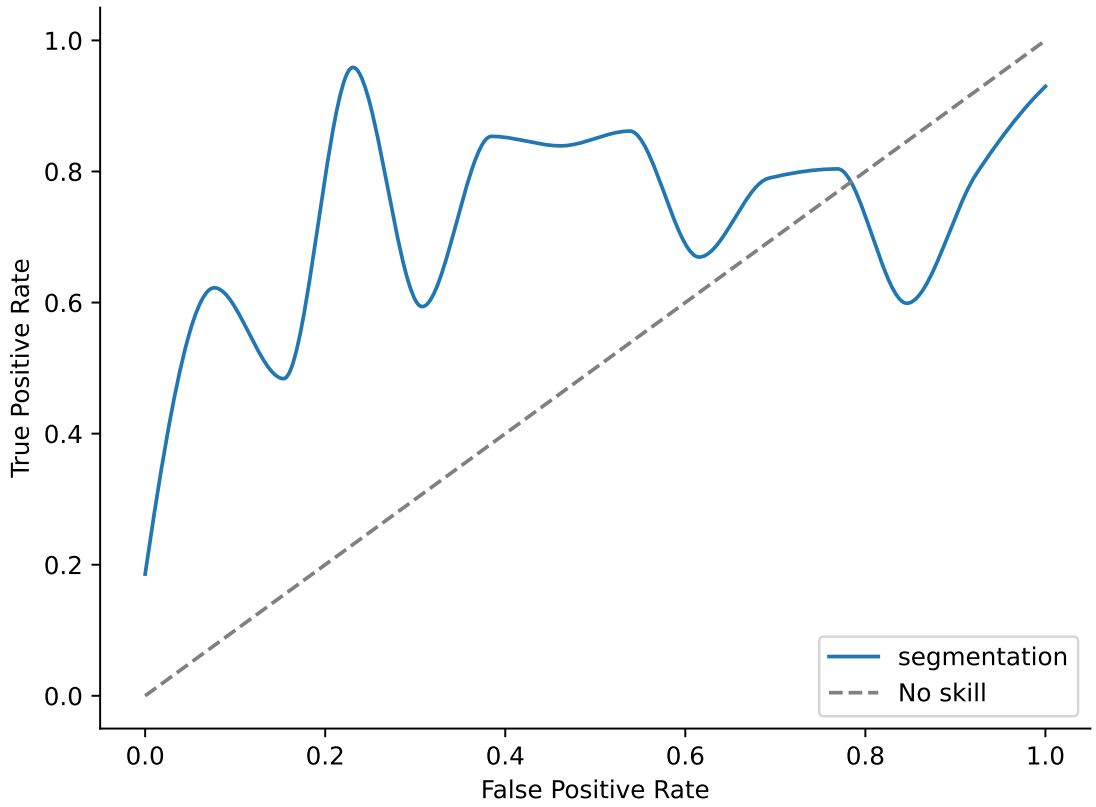

Figure 4.3: ROC curve on test data

dom, each from the training and the test set and computed the UMAP projections using the configuration $(10, 0.1, \text{cosine}, 42)$, representing the number of neighbors, minimum distance, similarity measure, and random seed, respectively. As shown in fig. 4.4, The embeddings from the training set show a high degree of coherence structurally, which indicates a well learned representation by the model internally.

In contrast to this the embeddings on the test set which are visualized in fig. 4.5 were more dispersed from the nnFormer baseline model but were a lot more compact and structured for the proposed model. This difference show that the inclusion of the SSP contribute to having a lot more stable and generalizable latent representation even on data that is previously unseen. The ability of the model to perform discriminatively and have consistent latent space representation is a major indicator of the strength of generalization in the performance of the model. These results further prove that the qualitative and the also the quantitative improvements that are gained through the anatomical regularization is very beneficial.

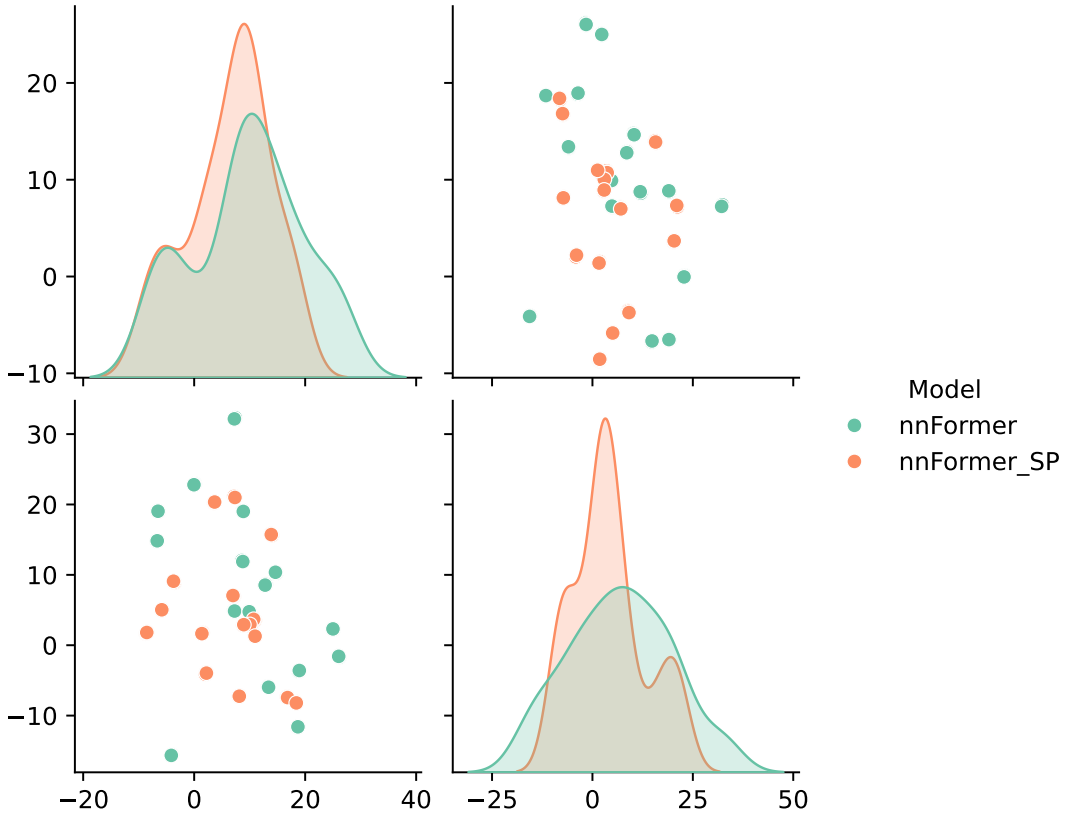

Figure 4.4: UMAP bottleneck feature representation on training data

4.6 Training with Varying Dataset Sizes

To evaluate the efficiency of the learning of the model even further, and also to understand the data requirements of the proposed architecture an additional series of experiments was conducted in which both the baseline nnFormer model and the nnFormer model improved with the SSP were trained progressively on increasing subsets of the training data. The sizes of the subsets included 10, 20, 30, 40, 50, and 60 patients. Each of the subsets was constructed randomly by taking a random sample from the full training dataset while ensuring consistent data preprocessing. For each of these subsets, the models were trained from scratch using the same hyperparameters, optimizer settings, and the augmentations which are described in the methodology chapter in detail. The goal of the experiments was to measure the performance gains as a functions of the dataset size and to asses how effectively the models could generalize rom the limited amount of training data. The performance of the models was measured on the full test set using four different key metrics: Dice coefficient, Recall, Precision and the IoU score. These results are summarized

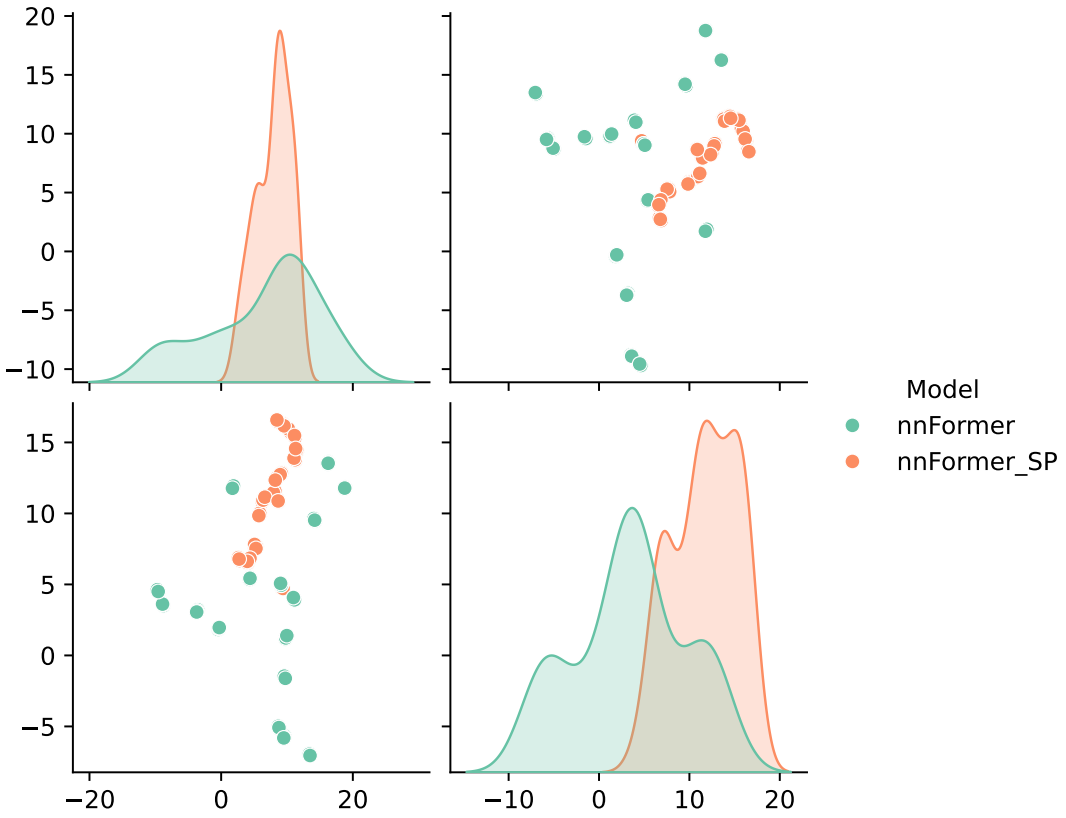

Figure 4.5: UMAP bottleneck feature representation on testing data

into a set of bar charts for each of the metrics, illustrated in fig. 4.6. These plots provide a very detailed view of the relationship between the size of the training set ans the performance of the model. As expected both the models show improvements in all the metrics as the size of the training set increases. However, the nnFormerSP consistently outperforms the baseline nnFormer at every stage of the training size. In the smaller subset size, the performance of the models are very inconsistent and random, the nnFormer baseline achieved a higher precision but the nnFormerSP achieves a higher score for each of the other 3 metrics, with having a huge difference in the values this alone proves the strength of incorporating shape priors in the training of the models and that it effectively compensates for the limited amount of training data. As the dataset size approaches patients,the performance of both the models are more consistent and stable and converge closer together but the nnFormerSP still retains a clear edge over the baseline in having better consistency and segmentation ability. These results further prove the advantage of using SSP in not only low-data settings but also in boosting the overall performance of the model

across varying training conditions. This experiment highlights the value of the shape-based regularization in DL pipelines, specifically in medical imaging especially when the availability of large amount of annotated data is difficult.

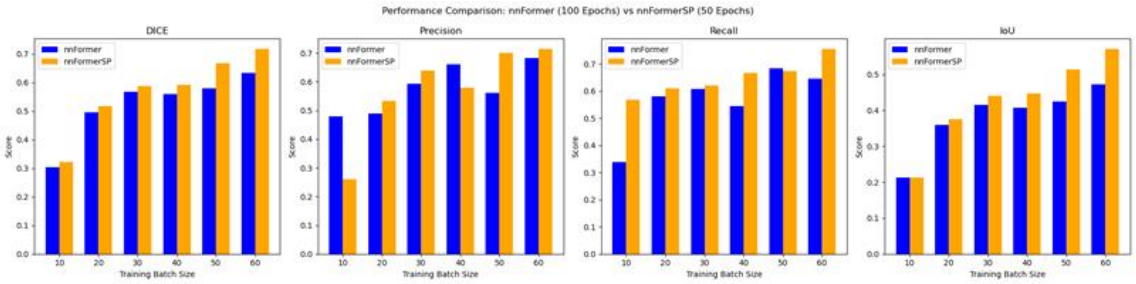

Figure 4.6: Evaluation metric variation based on dataset size

4.7 Conclusion of Results

The rsults of all the performed experiments demonstrate a very clear and a constant advantage of the proposed methodology over the baseline and other comparative approaches. From the synthetic noise resilience to the clinical accuracy on the real life patient data, the model proposed in the research not only delivers a higher performance in segmentation tasks but also introduces a reliability anatomically and strength of generalizablity through the use of SSPs. These attributes of a model are essential for the practical deployment of the segmentation tools in any real-world clinical settings.

Chapter 5

Conclusion

This research investigated the intersection of applications of DL with prior-based anatomical modeling in the context of MPI using SPECT. Specifically, this research shows the strength of integrating the transformer architecture, who are known for their capacity to learn long-range dependencies, with SSP which introduces a very powerful inductive bias that improves the accuracy of the segmentation, the robustness and the anatomical consistency. Through a very extensive set of experiments, both on the real-world patient data and also the synthetic phantoms the proposed hybrid model showed measurable improvements in all of the core metrics of segmentation when compared to the transformer baseline such as nnFormer and also the Swin-UNetR. Notably the proposed model excelled not only in high fidelity data settings but also in regimes where there is a limited amount of data available. Here, the SSPs played a very important role in guiding the training process and to reduce the overfitting of the model. These results demonstrate the advantage of using structured prior knowledge into data driven models especially in domains like medical imaging where the annotated data is not available in abundance, the anatomical plausibility is critical.

The methodology used in this study also proved to be extremely adaptable functioning very affectively across a range of SNR levels and also nuder a diverse amount of noise. The use of shape prior based phantoms in order to test the robustness to noise validated the generalization capacity of the the model even further and also its resilience to common imperfections that occur in the workflows of clinical imaging. Moreover, the analysis of the learned representations using UMAP embeddings indicated more structured and discriminative latent space for the method proposed

in this study, which suggests enhanced abstraction of features and a more in-depth understanding of the underlying distribution of anatomies. One of the major key finding is that even with a compact and modest dataset size of 50 to 60 labeled patients, the proposed model achieved a strong generalizational power to data that is unseen. This provides with an encouraging proof for future work in deploying such an architecture in real-world settings where te access to large amount of annotated data is most of the times impractical due to a lot of critical financial, ethical and operational limitations. The study confirms that reliable and standardized tools for segmentation of MPI SPECT are achievable without the necessity of a huge amount of data.

From the perspective of the clinics ,standardized and accurate segmentation can very significantly streamline the downstream tasks such as the functional quantification (such as Left ventricular ejection fraction (LVEF), EDV, ESV), rish mitigation and therap planning. The incorporation of transformer-based segmentation pipelines, augmented with the prior knowledge, lays a foundation for improving the diagnostic reproducibility and enhancing the decision making process of the clinics. Future work can build upon this foundation by incorporating dynamic priors exploring semi-supervised techniques, and validating the present approach across a number of different institutions and imaging modalities.

Acknowledgements

The research was supported by the project No. 2019-1.3.1-KK-2019-00011 financed by the National Research, Development and Innovation Fund of Hungary under the Establishment of Competence Centers, Development of Research Infrastructure Programme funding scheme. I would also like to thank my supervisor, and the open-source community for the successful completion of this project.

Bibliography

[1] I. Danad, P. G. Raijmakers, R. S. Driessen, *et al.*, "Comparison of coronary ct angiography, spect, pet, and hybrid imaging for diagnosis of ischemic heart disease determined by fractional flow reserve", *JAMA Cardiology*, 2nd vol., 10th no., pp. 1100–1107, Oct. 2017, ISSN: 2380-6583. DOI: 10.1001/jamacardio.2017.2471. eprint: https://jamanetwork.com/journals/jamacardiology/articlepdf/2648688/jamacardiology_danad_2017_oi_170038.pdf. [Online]. Available: https://doi.org/10.1001/jamacardio.2017.2471.

[2] P. Slomka, Y. Xu, D. Berman, and G. Germano, "Quantitative analysis of perfusion studies: Strengths and pitfalls", *Journal of Nuclear Cardiology*, 19th vol., 2nd no., pp. 338–346, 2012, ISSN: 1071-3581. DOI: https://doi.org/10.1007/s12350-011-9509-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071358123030908.

[3] E. V. Garcia, T. L. Faber, C. D. Cooke, R. D. Folks, J. Chen, and C. Santana, "The increasing role of quantification in clinical nuclear cardiology: The Emory approach", *Journal of Nuclear Cardiology*, 14th vol., SPEC. ISS.4th no., pp. 420–432, 2007, ISSN: 10713581. DOI: 10.1016/j.nuclcard.2007.06.009.

[4] Y. H. Liu, "Quantification of nuclear cardiac images: The Yale approach", *Journal of Nuclear Cardiology*, 14th vol., SPEC. ISS.4th no., pp. 483–491, 2007, ISSN: 10713581. DOI: 10.1016/j.nuclcard.2007.06.005.

[5] E. P. Ficaro, B. C. Lee, J. N. Kritzman, and J. R. Corbett, "Corridor4DM: The Michigan method for quantitative nuclear cardiology", *Journal of Nuclear Cardiology*, 14th vol., SPEC. ISS.4th no., pp. 455–465, 2007, ISSN: 10713581. DOI: 10.1016/j.nuclcard.2007.06.006.

[6] O. Z. Tolu-Akinnawo, F. Ezekwueme, O. Omolayo, S. Batheja, and T. Awoyemi, "Advancements in artificial intelligence in noninvasive cardiac imaging: A comprehensive review", *Clinical Cardiology*, 48th vol., 1st no., e70087, 2025.

[7] M. J. Henzlova and W. L. Duvall, "The future of spect mpi: Time and dose reduction", *Journal of nuclear cardiology*, 18th vol., 4th no., pp. 580–587, 2011.

[8] H. Xie, B. Zhou, X. Chen, *et al.*, "Transformer-based dual-domain network for few-view dedicated cardiac spect image reconstructions", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 163–172.

[9] H. Xie, W. Gan, W. Ji, *et al.*, "A generalizable 3d diffusion framework for low-dose and few-view cardiac spect", *arXiv preprint arXiv:2412.16573*, 2024.

[10] X. Chen, B. Zhou, X. Guo, *et al.*, "Dudocfnet: Dual-domain coarse-to-fine progressive network for simultaneous denoising, limited-view reconstruction, and attenuation correction of cardiac spect", *IEEE transactions on medical imaging*, 2024.

[11] H.-Y. Zhou, J. Guo, Y. Zhang, *et al.*, "Nnformer: Volumetric medical image segmentation via a 3d transformer", *IEEE transactions on image processing*, 32nd vol., pp. 4036–4045, 2023.

[12] E. P. Ficaro, B. C. Lee, J. N. Kritzman, and J. R. Corbett, "Corridor4dm: The michigan method for quantitative nuclear cardiology", *Journal of Nuclear Cardiology*, 14th vol., 4th no., pp. 455–465, 2007, Abstracts of Original Contributions, ASNC 2007, 12th Annual Scientific Session, ISSN: 1071-3581. DOI: https://doi.org/10.1016/j.nuclcard.2007.06.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071358107002942.

[13] E. V. Garcia, T. L. Faber, C. D. Cooke, R. D. Folks, J. Chen, and C. Santana, "The increasing role of quantification in clinical nuclear cardiology: The emory approach", *Journal of Nuclear Cardiology*, 14th vol., 4th no., pp. 420–432, 2007, Abstracts of Original Contributions, ASNC 2007, 12th Annual Scientific Session, ISSN: 1071-3581. DOI: https://doi.org/10.1016/j.nuclcard.

2007.06.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071358107002978.

[14] G. Germano, P. B. Kavanagh, P. J. Slomka, S. D. Van Kriekinge, G. Pollard, and D. S. Berman, "Quantitation in gated perfusion spect imaging: The cedars-sinai approach", *Journal of Nuclear Cardiology*, 14th vol., 4th no., pp. 433–454, 2007, Abstracts of Original Contributions, ASNC 2007, 12th Annual Scientific Session, ISSN: 1071-3581. DOI: https://doi.org/10.1016/j.nuclcard.2007.06.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071358107002966.

[15] Y.-H. Liu, "Quantification of nuclear cardiac images: The yale approach", *Journal of Nuclear Cardiology*, 14th vol., 4th no., pp. 483–491, 2007, Abstracts of Original Contributions, ASNC 2007, 12th Annual Scientific Session, ISSN: 1071-3581. DOI: https://doi.org/10.1016/j.nuclcard.2007.06.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071358107002930.

[16] M. Hosntalab, F. B. Mofrad, N. Monshizadeh, and M. Amoui, "Automatic left ventricle segmentation in volumetric SPECT data set by variational level set", *Int. J. Comput. Assist. Radiol. Surg.*, 7th vol., 6th no., pp. 837–843, 2012. DOI: 10.1007/S11548-012-0770-X. [Online]. Available: https://doi.org/10.1007/s11548-012-0770-x.

[17] T. Kohlberger, D. Cremers, M. Rousson, R. Ramaraj, and G. Funka-Lea, "4d shape priors for a level set segmentation of the left myocardium in spect sequences", in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, R. Larsen, M. Nielsen, and J. Sporring, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 92–100, ISBN: 978-3-540-44708-5.

[18] R. Yang, M. Mirmehdi, and D. Hall, "A charged contour model for cardiac spect segmentation", 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:2756798.

[19] R. Yang, M. Mirmehdi, X. Xie, and D. Hall, "Shape and appearance priors for level set-based left ventricle segmentation", *IET Computer Vision*, 7th vol., 3rd no., pp. 170–183, 2013. DOI: https://doi.org/10.1049/iet-cvi.2012.0081. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/

10.1049/iet-cvi.2012.0081. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cvi.2012.0081.

[20] C. Huang, X. Shan, Y. Lan, *et al.*, "A hybrid active contour segmentation method for myocardial d-spect images", *IEEE Access*, 6th vol., pp. 39 334–39 343, 2018. DOI: 10.1109/ACCESS.2018.2855060.

[21] X. Liu, M. Wegener, S. Polster, M. P. M. Jank, A. Roosen, and L. Frey, "Materials integration for printed zinc oxide thin-film transistors: Engineering of a fully-printed semiconductor/contact scheme", *Journal of Display Technology*, 12th vol., 3rd no., pp. 214–218, 2016. DOI: 10.1109/JDT.2015.2445378.

[22] J. Betancur and et al., "Automatic Valve Plane Localization in Myocardial Perfusion SPECT/CT by Machine Learning: Anatomic and Clinical Validation", *Journal of Nuclear Medicine*, 58th vol., 6th no., 2017, ISSN: 0161-5505. DOI: 10.2967/jnumed.116.179911.

[23] H. Arabi, A. AkhavanAllaf, A. Sanaat, I. Shiri, and H. Zaidi, "The promise of artificial intelligence and deep learning in pet and spect imaging", *Physica Medica*, 83rd vol., pp. 122–137, 2021, ISSN: 1120-1797. DOI: https://doi.org/10.1016/j.ejmp.2021.03.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1120179721001241.

[24] J. M. Wolterink, "Left ventricle segmentation in the era of deep learning", *Journal of Nuclear Cardiology*, 27th vol., 3rd no., pp. 988–991, 2020, ISSN: 1071-3581. DOI: https://doi.org/10.1007/s12350-019-01674-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071358123018809.

[25] T. Wang and et al., "A learning-based automatic segmentation and quantification method on left ventricle in gated myocardial perfusion spect imaging: A feasibility study", *Journal of Nuclear Cardiology*, 27th vol., pp. 976–987, 3 Jun. 2020, ISSN: 15326551. DOI: 10.1007/s12350-019-01594-2.

[26] H. Wen, Q. Wei, J.-L. Huang, *et al.*, "Analysis on spect myocardial perfusion imaging with a tool derived from dynamic programming to deep learning", *Optik*, 240th vol., p. 166 842, 2021, ISSN: 0030-4026. DOI: https://doi.

org/10.1016/j.ijleo.2021.166842. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402621005477.

[27] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS ser., (available on arXiv:1505.04597 [cs.CV]), vol. 9351, Springer, 2015, pp. 234–241. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a.

[28] S. Tang, J. Huang, G. Hung, *et al.*, "Dynamic programming-based automatic myocardial quantification from the gated spect myocardial perfusion imaging", in *The International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, Xi'an, China*, 2017, pp. 462–467.

[29] C. Zhao, S. Shi, Z. He, *et al.*, "Spatial-temporal v-net for automatic segmentation and quantification of right ventricle on gated myocardial perfusion spect images", *Medical Physics*, 50th vol., 12th no., pp. 7415–7426, 2023.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation", in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[31] Y. Zhang, F. Wang, H. Wu, *et al.*, "An automatic segmentation method with self-attention mechanism on left ventricle in gated pet/ct myocardial perfusion imaging", *Computer Methods and Programs in Biomedicine*, 229th vol., p. 107 267, 2023, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.107267. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722006484.

[32] F. Zhu, L. Li, J. Zhao, *et al.*, "A new method incorporating deep learning with shape priors for left ventricular segmentation in myocardial perfusion spect images", *Computers in Biology and Medicine*, 160th vol., p. 106 954, 2023, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2023.106954. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523004195.

[33] Y. Ni, D. Zhang, G. Ma, *et al.*, "A multiscale spatial transformer u-net for simultaneously automatic reorientation and segmentation of 3-d nuclear cardiac images", *IEEE Transactions on Radiation and Plasma Medical Sciences*, 8th vol., 6th no., pp. 632–645, 2024. DOI: 10.1109/TRPMS.2024.3382318.

[34] Á. I. Szundefinedcs, K. Horváth, K. Sólyomvári, Á. Zlehovszky, and B. Kári, "Self-supervised segmentation of myocardial perfusion imaging spect left ventricles", in *Proceedings of the 2023 10th International Conference on Bioinformatics Research and Applications*, ICBRA '23 ser., Barcelona, Spain: Association for Computing Machinery, 2024, pp. 206–211, ISBN: 9798400708152. DOI: 10.1145/3632047.3632078. [Online]. Available: https://doi.org/10.1145/3632047.3632078.

[35] H.-Y. Zhou, J. Guo, Y. Zhang, *et al.*, "Nnformer: Volumetric medical image segmentation via a 3d transformer", *Trans. Img. Proc.*, 32nd vol., pp. 4036–4045, Jan. 2023, ISSN: 1057-7149. DOI: 10.1109/TIP.2023.3293771. [Online]. Available: https://doi.org/10.1109/TIP.2023.3293771.

[36] K. Nagy, M. Tóth, P. Major, *et al.*, "Performance evaluation of the small-animal nanoscan pet/mri system", *Journal of Nuclear Medicine*, 54th vol., 10th no., pp. 1825–1832, 2013. DOI: 10.2967/jnumed.112.114785.

[37] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data", *IEEE Transactions on Medical Imaging*, 13th vol., 4th no., pp. 601–609, 1994. DOI: 10.1109/42.363108.

[38] A. Paszke and et al., "Pytorch: An imperative style, high-performance deep learning library", in Curran Associates, Inc., 2019, pp. 8024–8035.

[39] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy", *Nature*, 585th vol., 7825th no., pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python", *Journal of Machine Learning Research*, 12th vol., pp. 2825–2830, 2011.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980.

[42] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui, "4d xcat phantom for multimodality imaging research", *Medical Physics*, 37th vol., 9th no., pp. 4902–4915, 2010. DOI: https://doi.org/10.1118/1.3480985. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.3480985. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3480985.

[43] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images", in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds., Cham: Springer International Publishing, 2022, pp. 272–284, ISBN: 978-3-031-08999-2.

[44] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection", *Journal of Open Source Software*, 3rd vol., 29th no., p. 861, 2018. DOI: 10.21105/joss.00861. [Online]. Available: https://doi.org/10.21105/joss.00861.

List of Figures

List of Tables

List of Equations