Targeted Segmentation: Leveraging Localization with DAFT for Improved Medical Image Segmentation

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Samruddhi Shastri
2019111039
samruddhi.shastri@research.iiit.ac.in

International Institute of Information Technology Hyderabad
500 032, India

June 2024

To family and friends

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***Targeted Segmentation: Leveraging Localization with DAFT for Improved Medical Image Segmentation*** by *Samruddhi Shastri* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Jayanthi Sivaswamy

Acknowledgement

I would like to express my deepest gratitude to my research advisor, Dr. Jayanthi Sivaswamy. Her unwavering support, insightful guidance, and constructive feedback have helped me throughout my research journey. I am fortunate to benefit from her dedication to my academic success, and her personal support to help me work at my pace.

I would also like to extend my appreciation to my co-authors, Naren Akash R J and Lokesh Gautham. Their collaborative spirit, and willingness to share their knowledge have been essential to the development of this research. The insightful discussions and collaborative efforts we shared have significantly improved the quality of this work.

I am also grateful to my parents for their constant support and encouragement throughout my academic pursuits. Their belief in my abilities has been a source of strength and motivation, particularly during challenging moments. Their support played a crucial role in ensuring that I remained focused and on pace with the demands of this research journey.

I would like to express my heartfelt gratitude to my friends, Sasanka, Tanvi, Ahana, Chandana, and Pranjali. Their friendship provided a much-needed escape and a source of laughter and understanding. The moments we shared were invaluable in keeping me grounded and motivated.

My gratitude extends to my lab mates, Anirudh, Alphin, Abhishek, Rudransh, and Shreyas. Their presence in the lab fostered a collaborative environment. Their willingness to lend a hand with experiments, engage in discussions, and offer support during troubleshooting sessions proved invaluable throughout this research journey.

Finally, I would like to express my gratitude to all those who have directly or indirectly contributed to this research. Your support, in whatever form it may have taken, is deeply appreciated.

Abstract

Medical imaging plays a pivotal role in modern healthcare, providing clinicians with crucial insights into the human body's internal structures. However, extracting meaningful information from medical images, such as X-rays and Computed Tomography (CT) scans, remains a challenging task, particularly in the context of accurate segmentation. This thesis presents a novel two-stage Deep Learning (DL) pipeline designed to address the limitations of existing single-stage models and improve segmentation performance in two critical medical imaging tasks: pneumothorax segmentation in chest radiographs and multi-organ segmentation in abdominal CT scans.

The first stage of the proposed pipeline focuses on localizing target organs or lesions within the image. This initial localization stage utilizes a specialized module tailored to the specific organ/lesion and image type. This stage outputs a "localization map" highlighting the most probable regions where the target resides, guiding the next step. The second stage, fine-grained segmentation, precisely delineates the organ/lesion boundaries. This is achieved by combining UNet, known for its ability to capture both general and detailed features, with Dynamic Affine Feature-Map Transform (DAFT) modules that dynamically adjust information within the network. This combined approach leads to more accurate boundary delineation, meticulously outlining the exact borders of the target organ/lesion after roughly locating it in the first stage.

An application of the proposed pipeline focuses on pneumothorax segmentation, leveraging not only the image data but also the accompanying free-text radiology reports. By incorporating text-guided attention and DAFT, the pipeline produces low-dimensional region-localization maps, significantly reducing false positive predictions and improving segmentation accuracy. Extensive experiments on the CANDID-PTX dataset demonstrate the efficacy of the approach, achieving a Dice Similarity Coefficient (DSC) of 0.60 for positive cases and 0.052 False Positive Rate (FPR) for negative cases, with DSC ranging from 0.70 to 0.85 for medium and large pneumothoraces.

Another application of the proposed pipeline involves multi-organ segmentation in abdominal CT scans, where accurate delineation of organ boundaries is crucial for various medical tasks. The proposed Guided-nnUNet leverages spatial guidance from a ResNet-50-based localization map in the first stage, followed by DAFT-enhanced 3D U-Net (nn-UNet implementation). Evaluation on the AMOS and Beyond The Cranial Vault (BTCV) datasets demonstrates a significant improvement over baseline

models, with an average increase of 7% and 9% on the respective datasets. Moreover, Guided-nnUNet outperforms state-of-the-art (SOTA) methods, including MedNeXt, by 3.6% and 5.3% on the AMOS and BTCV datasets, respectively.

Overall, this thesis proposes a novel two-stage deep learning pipeline for medical image segmentation, demonstrating its effectiveness in handling a wide range of anatomical structures and image modalities (2D X-ray, 3D CT) for both single-organ (e.g., pneumothorax segmentation in chest radiographs) and multi-organ segmentation tasks (e.g., abdominal CT scans). This comprehensive approach offers significant advancements and contributes to improved medical image analysis, potentially leading to better healthcare outcomes.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
BTCV	Beyond The Cranial Vault
CBAM	Convolutional Block Attention Module
CE	Cross Entropy
CNN	Convolutional Neural Network
CT	Computed Tomography
DAFT	Dynamic Affine Feature-Map Transform
DL	Deep Learning
DSC	Dice Similarity Coefficient
FCN	Fully Convolutional Network
FP	False Positive
FPR	False Positive Rate
GPU	Graphics Processing Units
LCS	Label Conditioned Segmentation
MRI	Magnetic Resonance Imaging
PTX	Pneumothorax
ROI	Regions of Interest
SOTA	state-of-the-art
TN	True Negative
XAI	Explainable AI

Symbols

n	Customizable parameter that determining how finely the image/volume is subdivided
$\hat{\mathbf{L}}$	Low-dimensional localization map
O	Total number of organs/lesions
dim	Dimensionality of image
d	Segmentation decoder block index
α_d	Scaling parameter in DAFT module of d-th block
β_d	Shifting parameter in DAFT module of d-th block
F_d	Feature map of the d-th block of the decoder
F'_d	Modified feature map of the d-th block of the decoder
f	Function to calculate scaling parameter in DAFT module
g	Function to calculate shifting parameter in DAFT module
R_t	Text features from radiology reports
R_i	Visual features from radiology images
N_R	Number of channels in R_t
D_R	Dimension of embeddings in R_t
R_d	Feature map of d-th block of the ConTEXTualNet decoder
D_d	Depth of the feature map of the d-th block of the ConTEXTualNet decoder
$\mathbf{A}$	Pixel-wise attention map in ConTEXTualNet
Q	Upsampled image feature maps for language cross-attention
$\bar{Q}$	Flattened query vector derived from Q for language cross-attention
K	Key vector extracted from text features for language cross-attention
V	Value vector extracted from text features for language cross-attention
W_Q	Learnable weight that projects $\bar{Q}$ to a common space
W_K	Learnable weight that projects K to a common space

W_V	Learnable weight that projects V to a common space
d_k	Dimension of Q and K
Q^*	Modified Q highlighting the image regions that are relevant to the textual information
P_{pred}	Predicted segmentation mask
P_{gt}	Ground truth reference mask
$\cap$	Intersection
$\cup$	Union

Chapter 1

Introduction

Medical imaging has revolutionized healthcare, providing unparalleled insights into the human body. From X-rays to Magnetic Resonance Imaging (MRI), these technologies enable visualization of internal structures, aiding in diagnosis, treatment planning, and monitoring disease progression. However, extracting meaningful information from these images often requires further processing. In this context, medical image segmentation emerges as a powerful tool.

1.1 Introduction to Medical Image Segmentation

Medical image segmentation is a computer vision task that delineates Regions of Interest (ROI) within an image. By segmenting structures like organs, tissues, or lesions, clinicians gain a deeper understanding of the underlying anatomy and pathology. Inaccurate segmentation can lead to misdiagnosis, ineffective planning, and hampered disease monitoring.

The importance of medical image segmentation extends far beyond individual patient care. Improved segmentation algorithms have the potential to significantly impact healthcare systems economically and socially. By automating tedious manual segmentation tasks, these algorithms can free up valuable clinician time, leading to increased efficiency and potentially lower healthcare costs. Furthermore, accurate segmentation can enable the development of more precise diagnostic tools and personalized treatment plans, ultimately improving patient outcomes and quality of life.

1.2 Challenges in Medical Image Segmentation

Achieving accurate medical image segmentation encounters several challenges that go beyond the general difficulties of image segmentation. These challenges arise from the inherent complexities of medical images themselves and the crucial role segmentation plays in downstream clinical applications. They can be broadly categorized into three areas:

- **Image characteristics:** Medical images present inherent challenges for segmentation due to several factors. These include:
  - Variability: Images can vary significantly due to factors like acquisition device, patient positioning, and pathology itself.
  - Noise and artifacts: Images can be corrupted by noise from the imaging device or artifacts caused by patient movement or metal implants.
  - Low contrast: The boundaries between structures of interest may be faint or indistinct, making segmentation difficult.
- **Anatomical complexity:** Unlike everyday objects, the human body presents challenges like:
  - Varying organ shape and size: Organs and tissues can have complex, irregular shapes that vary significantly between individuals.
  - Overlapping structures: Organs and tissues can be closely packed, making it difficult to distinguish their boundaries.
  - Lesion heterogeneity: Disease manifestations can vary greatly in size, shape, and intensity, making it difficult to differentiate lesions from healthy tissue.
- **Data limitations:** Training robust segmentation algorithms requires a large amount of accurately labeled data, which can be a significant hurdle:
  - Annotation cost: Labeling medical images requires expertise from medical professionals, making it a time-consuming and expensive process.
  - Data privacy: Patient privacy regulations can restrict access to large datasets of medical images.
  - Class imbalance: In some cases, the disease of interest may be rare, leading to an imbalanced dataset where healthy tissue dominates. This can make it difficult for algorithms to learn to segment the less frequent disease class.

1.3 Deep Learning for Medical Image Segmentation

Medical image segmentation traditionally relied on manual techniques or rule-based algorithms, each with its own advantages and limitations. Common approaches include thresholding (Otsu's method and adaptive thresholding), edge detection (Sobel filter, Prewitt filter, and Canny edge detection), region-based techniques (region growing, seeded region growing, and watershed segmentation), and active contours (snakes and level sets) [1]. Thresholding segments images based on pixel intensity, while edge detection aims to identify boundaries between different regions. Region-based techniques group pixels based on shared characteristics like intensity or texture. Active contours use deformable models to

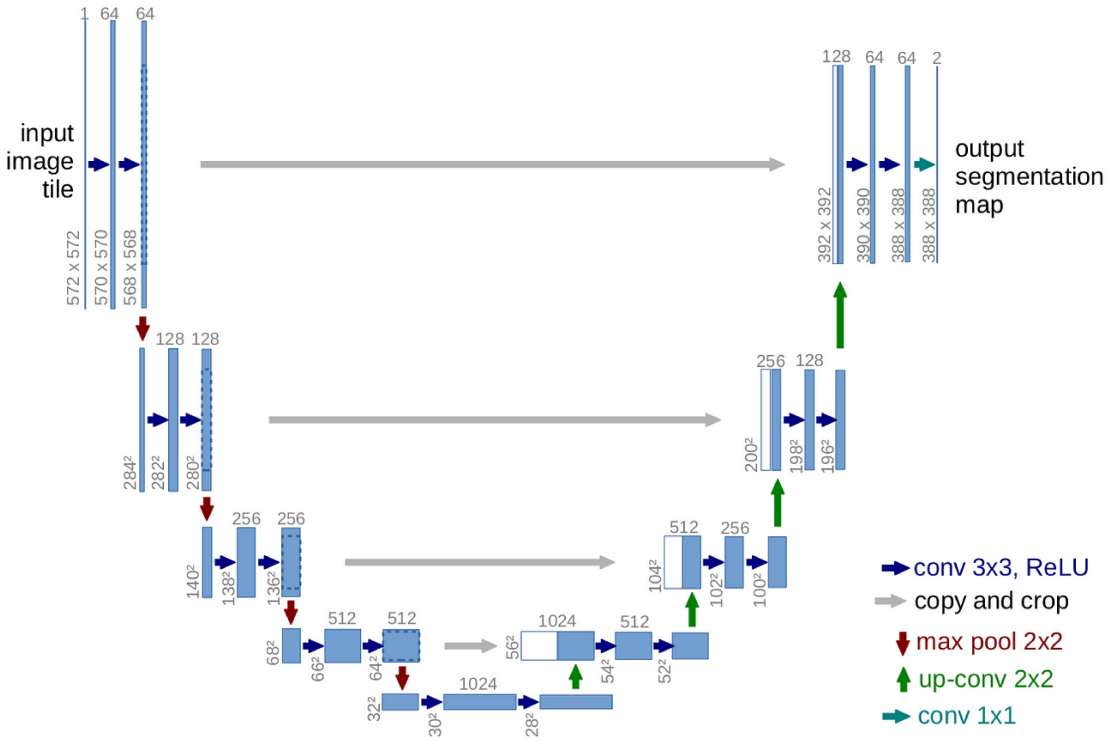


Figure 1.1 This diagram illustrates the U-Net architecture, a convolutional neural network commonly used for image segmentation tasks. It consists of a contracting path (encoder) that captures contextual information and a symmetric expanding path (decoder) that refines the features for segmentation. (source: www.towardsdatascience.com)

fit object boundaries iteratively [2]. While these methods have played a significant role, they require extensive preprocessing, parameter tuning, and can be sensitive to noise, intensity variation and spatial ambiguity within medical images [1].

The emergence of Artificial Intelligence (AI), particularly the field of DL, has offered promising solutions to these challenges. DL algorithms can learn intricate patterns and relationships within vast amounts of medical image data. This allows them to automatically segment images with high accuracy and efficiency. DL models can handle the inherent challenges of medical images by learning from large datasets, continuously improving their segmentation abilities over time. As a result, DL offers immense potential to revolutionize medical image segmentation. By improving accuracy, speed, and consistency, DL-powered segmentation can equip clinicians with the insights they need, to make more precise diagnoses, ultimately leading to better patient care.

One prominent approach within DL is Fully Convolutional Network (FCN). These architectures form the foundation for many segmentation techniques. FCNs utilize convolutional layers throughout the net-

work, allowing them to process entire images and produce pixel-wise segmentation masks. Examples include traditional FCN [3], SegNet [4], and DeepLab [5]. While efficient, FCNs can struggle with capturing long-range dependencies within the image. This limitation highlights the ongoing exploration of various DL techniques to address specific challenges in medical image segmentation.

Building upon the foundation of FCNs researchers have explored various techniques to address specific challenges in medical image segmentation. One such approach is Encoder-Decoder Networks. These architectures separate feature extraction and segmentation tasks. The encoder network extracts high-level features from the image, while the decoder network upsamples and refines these features to generate a detailed segmentation map. Popular examples include U-Net [6] (Figure 1.1), DeepMedic [7], nnU-Net [8], and H-DenseUNet [9]. While these models excel at producing detailed segmentation boundaries, they may require more training data compared to FCNs.

To further improve segmentation accuracy, particularly in scenarios with overlapping structures or class imbalances, Attention-based Networks have emerged. These models focus on informative regions within the image during segmentation. Examples include Transformers [10] adapted for medical images and Attention U-Net [11]. Additionally, the Convolutional Block Attention Module (CBAM) [12] can be integrated into various architectures to enhance focus on relevant features. However, incorporating attention mechanisms can result in increased model size and training time [13].

While, DL offers significant promise, existing single-stage models face inherent limitations. Unlike isolated objects, organs within the body exhibit well-defined spatial relationships. For instance, the liver consistently sits next to the right kidney. However, single-stage models often overlook these crucial spatial cues. This necessitates algorithms to reason about the anatomical structures and integrate this knowledge into the segmentation process, posing a challenge for single-stage architectures.

Furthermore, these models attempt to perform both coarse and fine-grained segmentation simultaneously. Coarse segmentation involves identifying general organ areas, while fine-grained segmentation focuses on accurately delineating boundaries. This dual task can create an information bottleneck within the model, leading to inaccurate boundaries, particularly when dealing with structures like pneumothorax, where collapsed lung regions can have subtle intensity variations and poorly defined edges [14].

Finally, single-stage frameworks may lack sufficient computational capacity, especially when handling multiple anatomical structures within a single image. Additionally, they can struggle with class imbalances that are common in medical imaging tasks. For example, in pneumothorax segmentation, the collapsed lung region might be a much smaller area compared to healthy lung tissue. Single-stage models might prioritize the dominant class (healthy lung) and compromise the segmentation accuracy

of the less frequent class (pneumothorax).

To address these issues, strategies such as Cascaded/Ensemble Models have been explored [15]. This approach combines multiple DL models with different architectures, leveraging their complementary strengths for potentially improved segmentation performance. However, such ensembles require careful selection and integration of individual models. Hybrid Models are also explored [9, 11], which combine elements from various categories, such as FCNs, encoder-decoder structures, and attention mechanisms. These models aim to benefit from the strengths of each technique, but their design and optimization can be complex. As research continues to evolve, these hybrid approaches hold significant promise for the future of medical image segmentation.

1.4 Thesis Focus

Recognizing the limitations of single-stage models, alternative approaches are crucial for achieving more robust medical image segmentation. Building on the established strengths of DL in this field, this thesis proposes a novel two-stage hybrid DL pipeline designed to improve segmentation performance. The research investigates distinct yet complementary challenges in two areas:

- **Pneumothorax Segmentation in Chest Radiographs:** Accurately identifying pneumothorax, a collapsed lung region, in chest X-rays is crucial for timely diagnosis and treatment. However, the subtle and variable appearance of pneumothorax can be challenging to detect solely based on image data. To address this limitation, this thesis introduces a novel two-stage approach that leverages additional information beyond the X-ray itself. This approach incorporates not only the image data but also the associated free-text radiology report, aiming to achieve more accurate and robust segmentation of pneumothorax compared to traditional methods.
- **Multi-Organ Segmentation in Abdominal CT Scans:** Precise segmentation of multiple organs within abdominal CT scans plays a vital role in various medical tasks. However, achieving accurate segmentation can be challenging due to the intricate spatial relationships between organs and the variations in their shapes and sizes. These complexities can lead to difficulties in correctly delineating boundaries, particularly when organs are close together or when some organs are significantly larger than the others. To address these challenges, we propose Guided-nnUNet, a two-stage segmentation framework that decomposes abdominal multi-organ segmentation into organ localization, followed by localization-guided fine segmentation.

This thesis comprehensively addresses the challenge of medical image segmentation. It demonstrates the ability of the proposed two-stage segmentation pipeline to handle a diverse range of anatomical structures and image characteristics. The research investigates segmentation in various scenarios, including single-organ segmentation in 2D X-ray images (e.g., pneumothorax) and multi-organ segmentation in

3D CT volume data (e.g., abdominal organs). By encompassing both 2D and 3D modalities along with single and multiple organs, this work highlights the versatility and generalizability of the approach. This paves the way for its potential application in various medical image segmentation tasks.

1.5 Summary of Contributions

This thesis addresses the challenges of medical image segmentation by proposing a novel two-stage DL pipeline that improves performance and overcomes limitations of existing single-stage models. The key contributions of this research are:

- Introducing a novel two-stage DL pipeline for improved segmentation performance.
- Application of the pipeline for pneumothorax segmentation from chest radiographs and associated free-text radiology reports.
- Cross attention-based rough localization of pneumothorax which leverages free-text radiology reports and using DAFT for fine segmentation based on rough localization information.
- Application of the pipeline for accurate multi-organ segmentation in abdominal CT scans.
- Using DAFT to fuse the organ localization information as spatial guidance to improve fine-grained segmentation.

1.6 Organisation of the Thesis

This thesis is organized into five chapters to provide a comprehensive exploration of medical image segmentation using a novel two-stage DL pipeline.

- **Chapter 1: Introduction** lays the groundwork by introducing the importance of medical imaging and the role of image segmentation. It then discusses the challenges of traditional segmentation techniques and highlights the potential of DL. Finally, it provides a brief overview of the research focus and the two-stage pipeline.
- **Chapter 2: Two-Stage Segmentation Pipeline** presents the core of the research - the novel two-stage DL pipeline. This chapter details the architecture of the pipeline, explaining the functionalities of each stage and how they work together to achieve improved segmentation performance.
- **Chapter 3: Pneumothorax Segmentation with Text-Guided Attention** focuses on the application of the two-stage pipeline for pneumothorax segmentation in chest radiographs. This chapter describes the specific modifications made to the pipeline for this task, including the use of text-guided attention to incorporate additional information beyond the X-ray image itself. The chapter then presents the methodology, results, and evaluation of the pipeline for pneumothorax segmentation.

- **Chapter 4: Abdominal Multi-Organ Segmentation with Guided-nnUNet** explores the application of the two-stage pipeline for segmenting multiple organs within abdominal CT scans. This chapter details the adaptations made to the pipeline for multi-organ segmentation, potentially using techniques like Guided-nnUNet. It then presents the methodology, results, and evaluation of the pipeline for multi-organ segmentation.
- **Chapter 5: Conclusion** summarizes the key findings of the thesis. It reiterates the contributions of the research, highlighting the development and effectiveness of the two-stage pipeline. The chapter also discusses limitations of this work and potential future directions for research in this area.

Chapter 2

Two-Stage Segmentation Pipeline

2.1 Introduction

Section 1.3 explored the limitations of single-stage DL models in medical image segmentation. These limitations stemmed from their inability to effectively leverage spatial relationships between organs, perform coarse and fine-grained segmentation simultaneously, and handle computational demands and class imbalances within medical images.

In response to these challenges, this chapter introduces a two-stage hybrid DL pipeline specifically designed to overcome the shortcomings identified in single-stage models. This pipeline addresses the need for anatomical reasoning and integration of spatial context by incorporating a dedicated localization stage. Furthermore, by separating coarse and fine-grained segmentation into distinct stages, the pipeline aims to improve overall segmentation accuracy, particularly for structures with subtle intensity variations and poorly defined edges.

This chapter delves into the details of this pipeline, outlining its two key stages: organ/lesion localization and fine-grained segmentation. We will explore the rationale behind this structure and how it addresses the limitations discussed previously. Additionally, we will examine the specific deep learning architectures employed in each stage and the reasoning behind their selection.

2.2 Workflow

The pipeline proposed in this chapter follows a two-stage workflow designed to address the limitations of single-stage models. This section details the two key stages and their interaction:

- **Localization:** The first stage of the pipeline focuses on identifying the presence and approximate location of the target organ or lesion within the medical image. An application-specific localization module is employed for this purpose. The specific architecture of this module will differ based on the targeted organ/lesion and the properties of the medical images for each application.

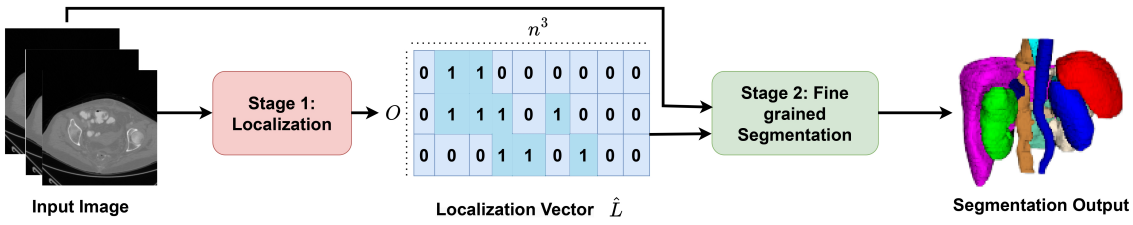


Figure 2.1 This diagram depicts the two-stage pipeline. Stage 1 performs localization, and Stage 2 utilizes this information for segmentation.

For instance, a Convolutional Neural Network (CNN) pre-trained on a large dataset of similar medical images might be suitable for some applications. In other cases, a lighter weight architecture or a different approach altogether might be more efficient, depending on the specific task. Regardless of the chosen architecture, the output of this stage is a localization map. This map highlights the most likely regions within the image that contain the organ or lesion of interest, providing valuable guidance for the subsequent segmentation stage.

- **Fine-grained segmentation:** The second stage of the pipeline tackles the precise delineation of the organ/lesion boundaries. It leverages a combination of UNet and DAFT modules. While UNet is employed in this work due to its well-established performance in medical image segmentation, it is important to note that DAFT can be integrated with other segmentation architectures as well. UNet is renowned for its ability to capture both high-level and low-level features from the medical image, a crucial aspect for achieving accurate segmentation. DAFT modules further enhance UNet's performance by dynamically modifying feature maps within the network.

2.3 Localization

The first stage of the pipeline employs a grid-based localization approach to identify the presence and approximate location of the target organ or lesion within the medical image. This section delves into the details of this method and explores the impact of grid resolution on its performance.

2.3.1 Grid Partitioning

Grid-based localization leverages a user-defined grid structure superimposed on the entire medical image (either 2D or 3D, depending on the application). This grid partitions the image into smaller sub-regions, essentially creating a coarse spatial map. The specific grid resolution is determined by a parameter denoted by n. A higher value of n translates to a finer grid with more numerous and smaller sub-regions, allowing for more precise localization of the target within the image. Conversely, a lower value of n results in a coarser grid with fewer and larger sub-regions, providing a less granular localization output.

2.3.2 Localization with the Grid

The entire medical image is fed into the chosen localization module (for example, CNN). This module then analyzes the image features and predicts a binary localization map $\hat{\mathbf{L}}$ with the same dimensions as the grid $O \times n^{dim}$, where O represents the number of target organs/lesions and dim represents the image dimensionality (2 for 2D images and 3 for 3D images). Each element within this binary map corresponds to a specific sub-region in the grid and signifies the presence of containing the target organ/lesion.

For instance, in a 2D image with a target organ and a grid resolution of $n = 4$, the localization module would output a binary map of size 1×16. Each of the 16 values in this map is either 1 (indicating the presence of the organ) or 0 (indicating the absence of the organ) for the corresponding sub-region (out of the 16 created by the 4×4 grid). Similarly, for a 3D image with a grid resolution of $n = 8$, the output binary map would have dimensions 1×512 ($8 \times 8 \times 8$), with each element being 1 if the corresponding sub-region contains the target lesion within the 3D volume and 0 otherwise.

2.3.3 Impact of Grid Resolution

The choice of grid resolution n plays a crucial role in the effectiveness of grid-based localization. A higher resolution grid (larger n) offers several advantages:

- Improved Localization Accuracy: With a finer grid, the localization module can provide more precise spatial information about the target's location within the image. This finer granularity can be particularly beneficial for smaller organs or lesions that might occupy only a portion of a sub-region in a coarser grid.
- Enhanced Differentiation: A higher resolution grid allows for a more nuanced differentiation between neighbouring sub-regions. This can be crucial for situations where multiple organs or lesions are located in close proximity within the image.

However, there are also drawbacks to consider with a higher grid resolution:

- Increased Computational Cost: Processing a finer grid with a larger number of sub-regions demands more computational resources from the localization module. This can potentially increase the training time for the entire pipeline.
- Potential for Overfitting: With a very high grid resolution, the localization module might struggle to generalize well during training, potentially leading to overfitting on the training data. In such cases, the model might not perform as well on unseen images with slightly different characteristics.

Therefore, selecting an optimal grid resolution requires careful consideration of the trade-off between localization accuracy, computational efficiency, and the risk of overfitting. In practice, the choice of n can be informed by factors such as the typical size of the target organ/lesion relative to the image dimensions, the desired level of localization precision, and the available computational resources.

2.4 Fine Grained Segmentation

Stage 2 of the pipeline tackles the task of precisely delineating the boundaries of the target organ or lesion localised in Stage 1. This stage leverages a combination of two deep learning architectures - U-Net and DAFT modules.

2.4.1 U-Net for Accurate Segmentation

The core architecture employed in this stage is the U-Net [6, 16]. Known for its efficiency in medical image segmentation tasks, U-Net excels at capturing both high-level semantic information (like the overall shape and location of organs) and low-level detailed features (like textures and boundaries) from the medical image. This ability to extract features across different scales is crucial for achieving accurate segmentation of organs and lesions with intricate structures and potentially subtle variations in intensity.

U-Net follows an encoder-decoder architecture. The encoder pathway progressively down-samples the input medical image, capturing high-level features essential for understanding the broader anatomical context. The decoder pathway then upsamples the encoded features and merges them with high-resolution features extracted earlier in the encoder path.

2.4.2 Incorporating Spatial Guidance with DAFT

While U-Net is a powerful segmentation architecture, it can sometimes struggle to incorporate spatial context into the segmentation process, particularly when dealing with complex anatomical structures [17]. To address this limitation, this work introduces DAFT modules within the U-Net framework.

DAFT [18] acts as a bridge, effectively fusing the rich visual features extracted from the raw medical image volume by U-Net with the coarse spatial guidance provided by the localization map generated in Stage 1. This localization map highlights the probable regions containing the target organ/lesion within the image. By incorporating this additional spatial information, DAFT empowers U-Net to focus its segmentation efforts on the most relevant image areas.

Its core lies in predicting scale (α_d) and shift (β_d) parameters for each feature map F_d of the decoder's d-th block:

$$F'_d = \alpha_d * F_d + \beta_d \tag{2.1}$$

where F'_d is the modified feature map. Scaling and shifting parameters are calculated as below:

$$\alpha_d = f(F_d, \hat{L}) \tag{2.2}$$

$$\beta_d = g(F_d, \hat{L}) \tag{2.3}$$

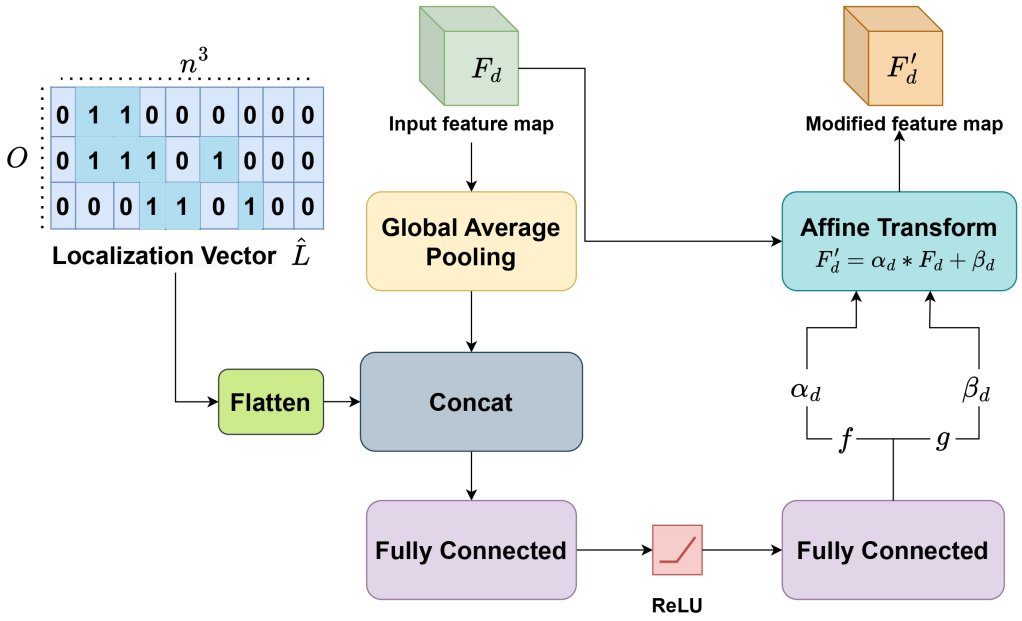


Figure 2.2 This diagram showcases the functionality of DAFT modules. DAFT analyzes the localization map and predicts adjustments (scaling and shift factors) for feature maps, ultimately refining them for more accurate segmentation.

Here, $\hat{L}$ denotes the grid-level localization, while f and g represent functions learned by a single auxiliary fully connected network.

Essentially, DAFT modulates the decoder's understanding of the image by *conditioning* it on the additional context provided by the previous stage. This conditioning is achieved through the predicted scaling and shifting parameters, effectively amplifying or suppressing specific features. Mathematically, these parameters control the extent to which the original features F_d are scaled and shifted, leading to a modified feature map F'_d incorporating weak localization information. By focusing on relevant image regions, DAFT guides the decoder towards a more accurate segmentation.

2.4.3 Combining U-Net and DAFT

The Stage 2 workflow can be summarized as follows:

- The original medical image and the localization map generated in Stage 1 are fed as inputs to the U-Net with DAFT modules.

- U-Net extracts feature maps from the medical image, capturing informative details at various scales.

- Within each decoder block, DAFT modules analyze the localization map and predict scaling and shift factors for the corresponding feature maps.
- These factors are applied to the feature maps, effectively modulating them based on the spatial guidance from the localization map.
- The modified feature maps are then processed through the decoder pathway of U-Net, allowing for precise localization and boundary delineation.
- The final output of Stage 2 is a binary segmentation mask that accurately delineates the boundaries of the target organ or lesion within the medical image.

2.5 Conclusion

This chapter details the two-stage pipeline that serves as the foundation for the tasks explored in further chapters. The rationale behind this structure lies in its ability to address the inherent limitations of single-stage models. Single-stage models often struggle to effectively reason anatomical relationships between organs, perform coarse and fine-grained segmentation simultaneously, and handle computational demands and class imbalances within medical images.

The two-stage approach separates localization from segmentation, allowing dedicated modules to excel at their respective tasks. Stage 1 identifies the target structure, while Stage 2 leverages this information to perform precise segmentation using a U-Net with DAFT modules. By employing dedicated modules for organ/lesion localization and segmentation, the pipeline promotes a more robust and efficient approach to medical image segmentation, laying the groundwork for particular applications investigated in the following chapters.

Chapter 3

Pneumothorax Segmentation with Text-Guided Attention

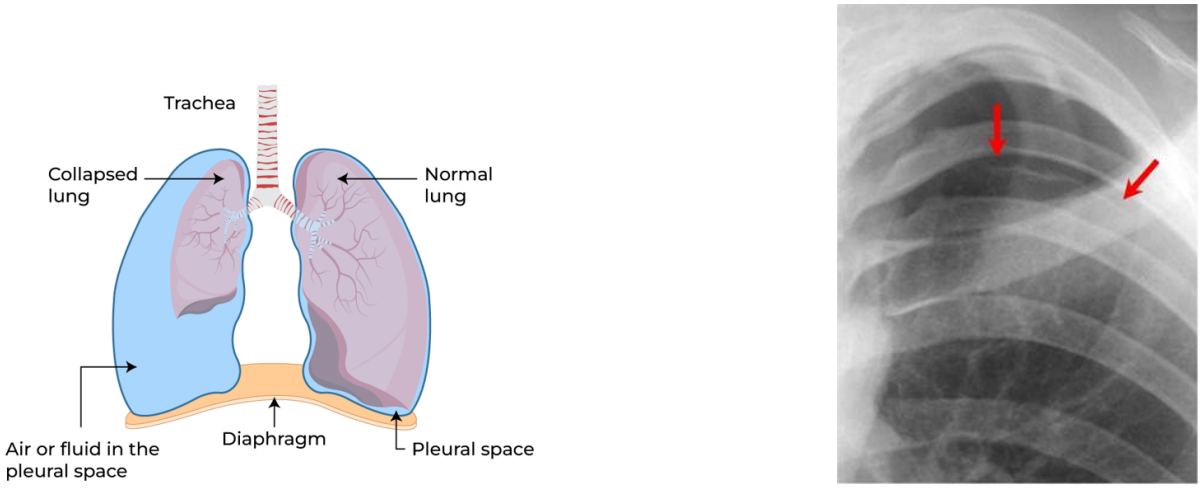
3.1 Introduction

Pneumothorax is a critical condition that occurs when air accumulates between the parietal and visceral pleura, leading to lung compression and hindering oxygen intake [19]. If left unnoticed, it can progressively worsen, potentially affecting other organs in the chest cavity (mediastinum), including the heart. Figure 3.1(a) shows the labelled diagram depicting pneumothorax.

Chest X-rays are often the first line of defence in diagnosing pneumothorax due to their cost-effectiveness, accessibility, and quick analysis time, making them especially valuable in severe situations. While any radiation exposure should be minimized, compared to CT scans, chest X-rays use much lower radiation doses. This is especially crucial for certain groups of patients who might need frequent monitoring or who have strict radiation limits to consider.

Although chest X-rays are frequently used for diagnosing pneumothorax, they do have limitations. The characteristic sign of a thin, sharp line representing the displaced lung can sometimes be faint and mistaken for normal anatomical structures or folds. This misinterpretation can occur due to the inherent limitations of two-dimensional imaging. Other structures, such as air-filled sacs (emphysematous bullae), skin folds, and even wrinkles in clothing, can also appear similar to pneumothorax on the X-ray. This can lead to potentially inappropriate pneumothorax management. Additionally, chest X-rays may not accurately assess the severity of pneumothorax, which is crucial for treatment decisions. Determining the size of the air leak and the extent of lung collapse is key to choosing the right intervention. However, accurately segmenting the collapsed lung region on X-rays can be challenging for clinicians due to overlapping anatomical structures like blood vessels and ribs. As illustrated in Figure 3.1(b), the pleural faint line can be barely discernible, making visual diagnosis challenging.

To overcome these hurdles, AI offers a powerful tool to complement the expertise of experienced clinicians. AI excels at pattern recognition, being able to discern subtle pneumothorax signatures that might elude human observation, such as faint pleural lines or minimal lung displacement. While ex-



(a) (b)

Figure 3.1 (a) A labelled diagram comparing a collapsed (pneumothorax-affected) lung to a normal lung (source: www.geeksforgeeks.org) (b) Chest X-ray section highlighting a collapsed lung (pneumothorax). Red arrows point to a thin white line, the visceral pleura. This single line is the key indicator of air trapped between the lung and chest wall, signifying a pneumothorax (source: www.learningradiology.com)

perienced clinicians are ultimately responsible for making final diagnoses, AI can be a valuable tool to enhance their expertise. By overcoming the limitations of chest radiographs, AI has the potential to significantly enhance patient outcomes while optimizing resource allocation in the management of pneumothorax.

Recent advancements in DL, such as multi-scale convolutional networks and U-Net architectures [6], have shown promising results in pneumothorax segmentation. These models can identify both subtle and prominent signs of a collapsed lung by extracting features at multiple scales. However, relying solely on image data has its limitations. One major drawback of image-based models is the lack of semantic information like the pneumothorax's location ('left apical region'), size ('small pneumothorax'), shape ('crescent-shaped'), and other features. Radiology reports provide valuable insights into these features. Without this rich textual data, image-based models may struggle with subtle characteristics or differentiating between pneumothorax and other lung pathologies.

To address this limitation, recent studies use multimodal approaches that combine chest X-rays and radiology reports [20, 21]. By incorporating textual descriptions' rich semantic information, these models can potentially achieve more accurate and robust pneumothorax segmentation. However, integrating image and text data presents its own set of challenges. The differences in representation and structure between images and text must be efficiently bridged for effective information fusion. Accurately matching corresponding image regions with relevant textual descriptions is also a crucial step in guiding the

model and improving its understanding of the findings.

Although existing models like ConVIRT [20] and GLoRIA [21] achieved promising results in downstream classification tasks by pre-training vision models on image-report pairs, they did not integrate text to guide image analysis. Other approaches [22] employed image and text encoders to identify instances of pneumonia and placed bounding boxes around them but did not provide pixel-level segmentation. Additionally, LAVT [23] was developed for referring-image segmentation of household items rather than medical images.

Although LVIT [24] achieved promising results with chest X-ray segmentation for COVID-19 patients, it still utilized synthesized text instead of real free-form radiology reports. Synthetic text lacks the nuance and richness of real reports, potentially hindering the model's ability to handle complex cases or variations in clinical language. The performance of another model, CPAM [25], relies heavily on the quality and specificity of the textual descriptions, giving misleading segmentation results in inaccurate or ambiguous reports. Despite ConTEXTualNet's [26] success in identifying pneumothorax with the aid of free-text reports, its exclusive training on positive samples creates a potential bias towards pneumothorax cases.

Building upon the two-stage segmentation pipeline detailed in Chapter 2, this chapter explores its application for pneumothorax segmentation in chest X-rays. By leveraging this pipeline alongside free-text radiology reports, we aim to overcome the limitations of existing models and achieve improved accuracy in pneumothorax segmentation.

3.2 Proposed Method

3.2.1 Workflow

In the two stage network (Figure 3.2), the first stage - Report-Guided Region Localization, aims to identify potential regions containing a pneumothorax. It starts by processing the free-text report using a language encoder to extract relevant text features (denoted as R_t). These text features are then combined with features (denoted as R_i) extracted from the chest X-ray image using ConTEXTualNet [26]. The resulting output is further analyzed by dividing it into quadrants and applying max-pooling to each quadrant. This step allows the network to infer the presence of pneumothorax in different image regions. Finally, this stage generates a localization map $\hat{\mathbf{L}}$, highlighting the most likely areas containing pneumothorax.

The second stage, Region-Aware Pneumothorax Segmentation, leverages $\hat{\mathbf{L}}$ generated in the first stage, to guide the segmentation process by highlighting potential pneumothorax regions within the chest X-ray. The 2D U-Net with DAFT modules, as detailed in Chapter 2, utilizes this localization

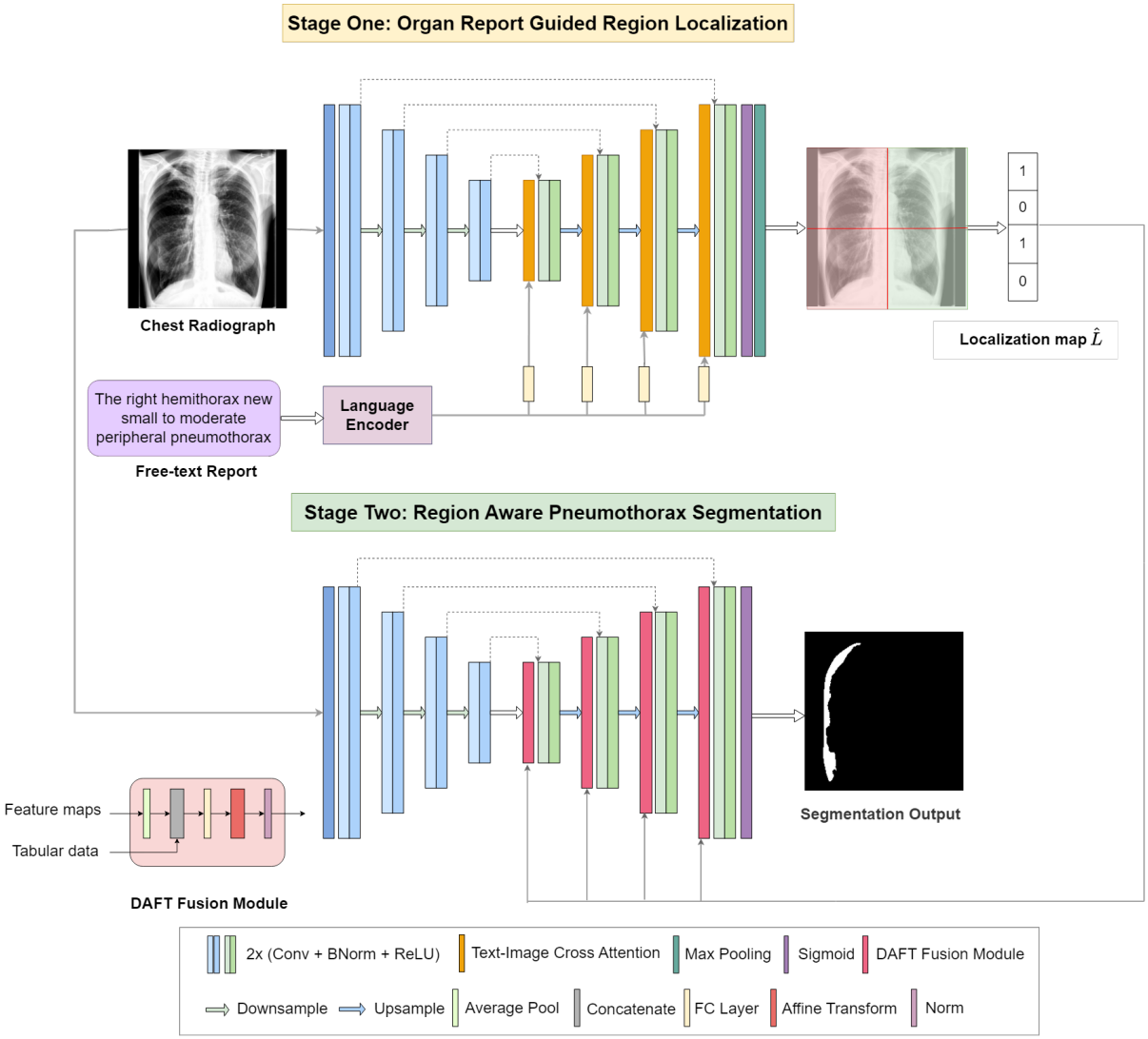


Figure 3.2 A schematic illustration of the proposed approach. Given a chest radiograph I and its corresponding free-text radiology report R, we first obtain a region localization map $\hat{\mathbf{L}}$, leveraging text-guided attention. In the subsequent stage, we modulate the feature maps of the segmentation network at multiple scales using $\hat{\mathbf{L}}$ to accurately segment pneumothorax.

information to refine its focus on these specific areas. This targeted approach allows for more precise delineation of the pneumothorax boundaries, resulting in a finely-segmented mask as the final output.

3.2.2 Report-Guided Region Localization

The first stage leverages a specialized model, ConTEXTualNet, to process the free-text radiology report. This model comprises cross-attention layers that combine information from both the input image and the accompanying text. To achieve this, a language encoder is employed to extract a set of

text features, labeled as $R_t \in \mathbb{R}^{N_R \times D_R}$, from the report. Here, N_R represents the number of channels, and D_R denotes the dimension of the embeddings. Later, a fully connected layer projects R_t into $R_d \in \mathbb{R}^{N_R \times D_d}$, where D_d indicates the depth of the feature map of the d-th block of the decoder. The text embeddings, enriched with knowledge about pneumothorax, interact with visual features to create a pixel-wise attention map $\mathbf{A}$, which highlights relevant image regions based on the textual information. To ensure compatibility between the two types of features, the query vectors derived from upsampled image features undergo projections alongside the key and value vectors from the text embeddings. The resulting contextualized feature maps are merged with encoder feature maps using skip connections in each decoder layer. The detailed implementation of cross attention is explained in Section 3.2.2.1.

After generating the attention map using the ConTEXTualNet, the next step is to obtain a rough localization map at the output of the decoder. To achieve this, the method is guided by semi-quantitative visual assessment methods [27] applied to chest radiographs. This involves dividing the image into four quadrants, two along each dimension ($n = 2$). Using a max-pooling layer on the final decoder feature map, the presence or absence of pneumothorax in each region (quadrant) can be inferred. This information is then summarized in a low-dimensional localization map $\hat{\mathbf{L}} \in 1 \times \{0, 1\}^4$.

3.2.2.1 Language Cross Attention

To generate the pixel-wise attention map $\mathbf{A}$, the ConTEXTualNet [26] uses multihead cross-attention mechanisms [13] that involve flattened query vectors $\bar{Q}$ from the upsampled feature map Q, and projecting key K and value V vectors from the text embeddings. This projection is carried out with the help of weights W_Q, W_K, and W_V, which facilitate the alignment of these vectors in the same space. The mathematical formula to calculate $\mathbf{A}$ is given in Equation 3.1.

$$A = \text{softmax}\left(\frac{\bar{Q}W_Q(KW_K)^T}{\sqrt{d_k}}\right)VW_V \tag{3.1}$$

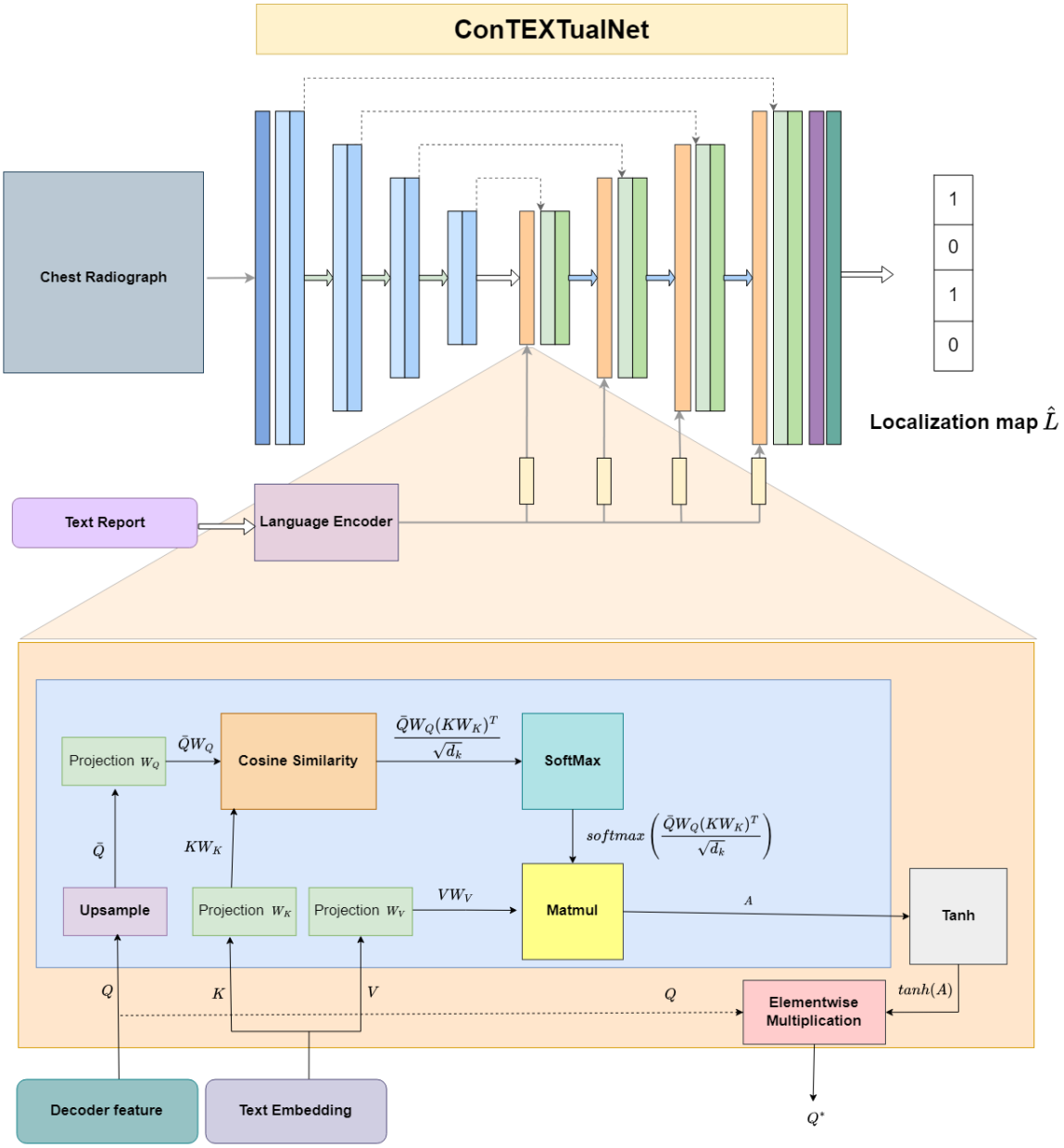
where, d_k represents dimension of Q and K. After generating the attention map $\mathbf{A}$, it is normalized using the Tanh activation function, which confines its values between -1 and 1. The normalized map is then applied pixel-wise to the query feature map Q using Equation 3.2, resulting in Q^* which helps to highlight the image regions that are relevant to the textual information.

$$Q^* = \tanh(\mathbf{A}) * Q \tag{3.2}$$

Figure 3.3 illustrates the complete workflow of Language Cross Attention module described above.

3.2.3 Region-Aware Pneumothorax Segmentation

Leveraging the information from the first stage, this final stage meticulously segments the pneumothorax region within the chest X-ray. It employs UNet architecture fused with DAFT modules specifically



$$A = softmax\left(\frac{\bar{Q}W_Q(KW_K)^T}{\sqrt{d_k}}\right)VW_V \qquad Q^* = tanh(A) * Q$$

Figure 3.3 This illustration depicts the workflow of the Language Cross Attention module. It takes decoder features (visual information) and text embedding (textual information) as input. The output, denoted by Q^*, highlights the image regions most relevant to the provided text.

designed to incorporate additional contextual information. The technical details of this architecture are presented in Section 2.4.

3.3 Dataset and Experimental details

3.3.1 Dataset

CANDID-PTX is a public dataset that contains 19,237 X-rays and corresponding anonymized free-text radiology reports from adult patients (aged 16 and above) collected at Dunedin Hospital, New Zealand. Each image has precise annotations marking pneumothoraces (collapsed lungs), acute rib fractures, and intercostal chest tubes. As part of this work, we have only used the annotations for pneumothorax. The dataset has annotations for 3,561 cases of collapsed lungs (pneumothorax) across the dataset. With a 1:5 positive-to-negative case ratio, the dataset provides a balanced mix of both healthy and pneumothorax-affected lungs, promoting robust training of algorithms.

3.3.2 Augmentations

To diversify the training data and improve model robustness, we employed a variety of image augmentation techniques inspired by previous pneumothorax segmentation research.

- **Color and Light Adjustments:** Randomly applying either contrast, gamma, or brightness changes 30% of the time.
- **Geometric Distortions:** Randomly applying elastic, grid, or optical distortions 30% of the time.
- **Affine Transformations:** Randomly scaling, rotating, and shifting images.

Horizontal flipping was discovered to disrupt the consistency between images and the accompanying text data. Therefore, it was excluded from all experiments. All augmentations were efficiently implemented using the Albumentations library [28].

3.3.3 Implementation Details

The chest radiographs were resized to a standard 224×224 dimension. Our experimental setup involved a stratified (by size of pneumothorax) five-fold cross-validation. Each fold included a designated testing set, while the remaining data was split into 75% for training and 25% for validation. The training process utilized the AdamW optimizer with an initial learning rate and weight decay of $1e{-}4$, implemented in PyTorch. The 2D U-Net with a pre-trained ResNet-50 backbone was utilized for segmentation. A frozen pre-trained T5-Large model was used to extract language embeddings from free-text reports. The two stages were trained sequentially and employed the weighted combination of

Methods	**PTX-Positive ↑**	**Small PTX ↑**	**Medium PTX ↑**	**Large PTX ↑**	**PTX-Negative ↓**
U-Net	0.550 ± 0.019	0.398 ± 0.015	0.635 ± 0.033	0.791 ± 0.046	0.738 ± 0.081
LViT	0.549 ± 0.010	0.378 ± 0.020	0.635 ± 0.012	0.798 ± 0.033	0.453 ± 0.095
CPAM	0.507 ± 0.031	0.343 ± 0.028	0.598 ± 0.038	0.751 ± 0.041	0.295 ± 0.201
ConTEXTualNet	0.566 ± 0.008	0.403 ± 0.020	0.657 ± 0.018	0.806 ± 0.029	$\mathbf{0.037} \pm 0.011$
Proposed	$\mathbf{0.601} \pm 0.013$	$\mathbf{0.429} \pm 0.024$	$\mathbf{0.697} \pm 0.011$	$\mathbf{0.851} \pm 0.017$	0.052 ± 0.011

Table 3.1 Comparison of pneumothorax (PTX) segmentation on the CANDID-PTX dataset. Five-fold average DSC (positive images) and false positive rate (negative images), with standard deviation, are listed for baseline and SOTA.

binary cross-entropy and Dice loss. The experiments were conducted on two NVIDIA GeForce RTX-2080Ti Graphics Processing Units (GPU). The training was limited to a maximum of 100 epochs with a batch size of 8.

3.3.4 Evaluation Metrics

The performance of the segmentation methods was evaluated using the DSC, given by

$$\text{DSC} = \frac{2 \times |P_{\text{pred}} \cap P_{\text{gt}}|}{|P_{\text{pred}}| \cup |P_{\text{gt}}|} \tag{3.3}$$

where P_{pred} and P_{gt} are the predicted segmentation mask and ground truth reference mask, respectively. The positive cases were subdivided into three classes based on the size of the pneumothorax in the image: small, medium and large, determined by thresholding. These thresholds were chosen based on the frequency histogram of the collapsed lung area on the chest radiograph. The performance was evaluated for each class.

For negative cases, the FPR was also calculated, focusing on correctly identifying the absence of pneumothorax. FPR is defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP + TN}} \tag{3.4}$$

where False Positive (FP) and True Negative (TN) represent the number of false positive predictions and true negative cases, respectively. A lower FPR signifies a better performance, indicating fewer false alarms or misclassifications of negative images as positive.

3.4 Results and Discussion

3.4.1 Comparison with baseline U-Net model

This section compares our proposed solution, which leverages a two-stage approach to a baseline U-Net model. U-Net relies solely on the chest X-ray images for pneumothorax segmentation without any prior localization step.

As shown in Table 3.1, U-Net achieves a mean DSC of 0.550 for positive cases (cases with pneumothorax). It also has an FPR of 0.738, indicating a high number of false positives. Our proposed significantly outperforms the U-Net baseline. In positive cases, our method achieves a 9.3% improvement in DSC, indicating a more accurate segmentation of the pneumothorax region. We can also see improvement in detection of false positives due to decrement in FPR by 93%. The improvement in DSC is consistent across varying pneumothorax sizes, highlighting the robustness of our approach. Our pipeline achieves a minimum improvement of 7.6% and a maximum improvement of 9.8% over the U-Net across different sizes.

3.4.2 Comparison with state-of-the-art methods

This section delves into the segmentation performance of our proposed two-stage approach compared to existing methods. Table 3.1 compares the segmentation performance of our proposed solution with baseline U-Net and the SOTA medical vision-language frameworks - LViT [24], ConTEXTualNet [26] and CPAM [25]. The mean DSC and FPR and the standard deviation across five folds are provided in Table 3.1.

In positive cases, our proposed solution outperforms LViT by 9.5%, ConTEXTualNet by 6.2%, and CPAM by 18.5%. The superior performance is sustained across varying sizes of the pneumothorax, as seen from the best performance (shown in bold font) being achieved by our method for all sizes. Specifically, our approach yields a min/max boost of 6.6%, to 13.5% over LViT; 13.3% to 25.1% over CPAM; and 5.6% to 6.5% over ConTEXTualNet.

The improvement in the performance of our method in positive cases over all these three methods can be attributed to the two-stage design. Additionally, the boost in segmentation performance for medium and large pneumothoraces directly addresses a critical need in clinical practice [29]. Our two-stage approach aligns with the priorities of human experts and provides a valuable tool for more efficient and accurate pneumothorax diagnosis.

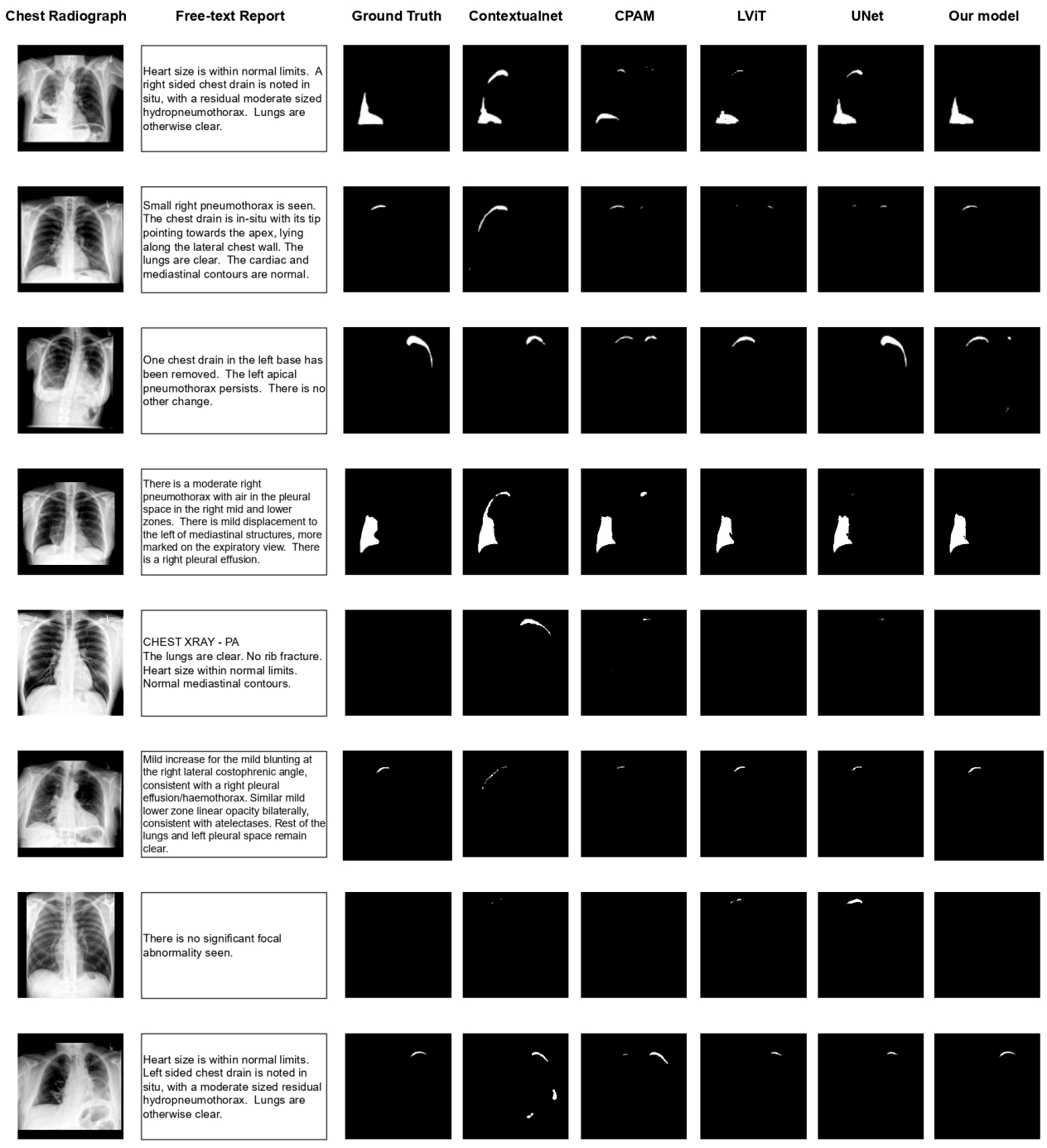


Figure 3.4 Qualitative results of pneumothorax segmentation by different methods.

Variants	Architecture	Text	PTX-Positive ↑	PTX-Negative ↓
V1	Unet	-	0.590 ± 0.010	0.686 ± 0.005
V2	-	Variable	0.591 ± 0.011	0.814 ± 0.005
V3	ConTEXTualNet	Constant	0.591 ± 0.013	0.786 ± 0.072
V4	ConTEXTualNet	Variable	$\mathbf{0.601 \pm 0.013}$	$\mathbf{0.052 \pm 0.011}$

Table 3.2 Comparison of stage 1 variants of our model for pneumothorax (PTX) segmentation on the CANDID-PTX dataset. Five-fold average Dice similarity coefficient (positive images) and false positive rate (negative images), with standard deviation.

3.4.3 Qualitative Analysis

Figure 3.4 provides a visual comparison of segmentation results using our method and four other approaches. It includes sample chest X-ray images, corresponding radiology reports, and the generated segmentation masks.

By analyzing the segmentation outputs, we can observe the strengths of our proposed method. Notably, U-Net and LViT suffer from under-segmentation in some cases and gives incorrect segmentation output (for example showing right pneumothorax instead of left). This suggests that these models might miss critical areas of the collapsed lung region. Conversely, ContextualNet, which incorporates text reports, exhibits over-segmentation. This indicates that it might be including irrelevant image regions in the segmentation mask.

In contrast, our method demonstrates a more balanced segmentation performance. This visual comparison highlights the potential of our approach to overcome limitations observed in other existing methods.

3.5 Ablation Studies

3.5.1 Stage 1

These experiments focus on evaluating the contribution of different design choices in Stage 1, responsible for generating the localization vector. The results are shown in Table 3.2:

3.5.1.1 Using U-Net for Stage 1 instead of ConTEXTualNet

This experiment replaces the original Stage 1 architecture (ConTEXTualNet, referred to as V4) with a simpler U-Net model (V1). We see that V4 outperforms V1 for positive as well as negative cases. It suggests that the text guidance provided by ConTEXTualNet is beneficial for capturing crucial details for localization as well as reducing false positives, in our pipeline.

Variants	DAFT-placement (B/D/E+D)	n	Mask or $\hat{\mathbf{L}}$	Localization	PTX-Positive ↑	PTX-Negative ↓
V1	B	4	$\hat{L}$	Variable	0.598 ± 0.009	0.053 ± 0.013
V2	D	4	$\hat{L}$	Variable	$\mathbf{0.601 \pm 0.013}$	$\mathbf{0.052 \pm 0.011}$
V3	E+D	4	$\hat{L}$	Variable	0.581 ± 0.012	0.082 ± 0.018
V4	B	6	$\hat{L}$	Variable	0.600 ± 0.013	0.067 ± 0.018
V5	-	-	Mask	-	0.567 ± 0.008	0.076 ± 0.010
V6	B	4	$\hat{L}$	Constant	0.587 ± 0.010	0.698 ± 0.086

Table 3.3 Comparison of stage 2 variants of our model for pneumothorax (PTX) segmentation on the CANDID-PTX dataset. Five-fold average Dice similarity coefficient (positive images) and false positive rate (negative images), with standard deviation.

3.5.1.2 Extracting Vector from Text instead of Stage 1

This ablation removes Stage 1 entirely (V2) and extracts an embedding vector directly from the text report. Since the baseline model (V4, with Stage 1) outperforms V2 (no Stage 1), it strongly suggests that the dedicated processing in Stage 1 plays a crucial role. Stage 1 likely generates a more informative localization vector that effectively guides Stage 2 for accurate segmentation compared to a simple embedding from the raw text.

3.5.1.3 Giving Constant Text in Stage 1 ConTEXTualNet

This variation (V3) explores the model's dependence on textual cues for localization. In V3, Stage 1 (ConTEXTualNet) receives a constant and generic text input that doesn't mention pneumothorax ('The heart size is normal. The lungs are clear. There is a focal eventration of the right hemidiaphragm. No mediastinal abnormality is seen'). The performance drop seen from V4 (actual reports) to V3 as shown in Table 3.2 strongly supports the hypothesis that textual information plays a crucial role in Stage 1.

3.5.2 Stage 2

These experiments focus on evaluating the impact of specific components within the Stage 2 segmentation framework. The results are shown in Table 3.3:

3.5.2.1 DAFT Placement

DAFT is a key component in Stage 2 that modulates the decoder's understanding of the image based on the localization information. This experiment explores the effectiveness of DAFT placement. We test three scenarios:

- DAFT only in the bottleneck layer (B) of the encoder-decoder architecture (V1)

- DAFT only in the decoder layers (D) (V2)
- DAFT included in both encoder (E) and decoder (D) layers (V3)

As shown in Table 3.3, V2 achieves superior performance compared to V1 and V3 which suggests that decoder-centric modulation might be more effective. The encoder extracts general image features, while the decoder refines them for segmentation. Placing DAFT in the decoder allows it to directly target this refinement process based on the localization information, leading to more accurate segmentation.

3.5.2.2 Varying Grid Resolution

This experiment investigates the influence of the grid resolution n used in the Stage 2 segmentation process. The grid resolution defines the granularity of the segmentation output. By testing different grid resolutions, we can find the optimal balance between capturing details and computational efficiency. We have experimented with $n = 4$ (V1) and $n = 6$ (V4).

By examining the performance metrics of V1 and V4 on the segmentation task, shown in Table 3.3, we see that V4 shows improvement in performance for positive cases. This suggests that capturing finer details can be beneficial for accurate segmentation of these positive cases. However, V4 also exhibits a decrease in performance for cases where there is no pneumothorax (negative cases). A possible reason for this is that higher resolution grid might have led to the model overfitting to the positive training examples with pneumothorax. This could result in the model being too focused on specific features associated with pneumothorax, potentially causing it to misclassify some negative cases (no pneumothorax) that have subtle differences from the training data.

3.5.2.3 Appending Mask in Stage 2 Bottleneck instead of Localization Vector

In this experiment, instead of feeding the localization vector into the bottleneck layer of Stage 2 (V1), we directly append the segmentation mask obtained from Stage 1 (V5). This allows us to assess if directly providing the segmentation mask as guidance is more effective than the learned localization vector. The results are shown in Table 3.3

Since V1 outperforms V5, it suggests that the model can learn more informative representations from the localization vector through Stage 1. This learned localization vector might be more effective in guiding the segmentation process in Stage 2 compared to a potentially noisy or inaccurate segmentation mask directly obtained from Stage 1.

3.5.2.4 Constant Localization

Here, we provide a constant localization vector (all ones) (V6) to Stage 2. This helps us understand how the model performs when the localization information from Stage 1 is not informative or missing.

The results are shown in Table 3.3.

Since V6 exhibits a decrease in segmentation accuracy compared to the normal performance, it strongly suggests that the model heavily relies on the localization information from Stage 1. The learned localization vector helps the model focus on the relevant image regions during segmentation, leading to more accurate results.

3.6 Conclusion

This chapter presents an application of the two-stage pipeline introduced in Chapter 2 for accurate segmentation of pneumothorax in chest radiographs. The free-text radiology reports have been used only in [26] to aid segmentation. The presented results highlight the efficacy of our proposed approach for pneumothorax segmentation, as well as provide insights into the impact of different design choices within each stage. Our method outperforms baseline models like U-Net and SOTA medical vision-language frameworks like LViT, ConTEXTualNet, and CPAM in terms of DSC for positive cases and FPR for negative cases.

In the ablation studies, we systematically analyzed the contributions of various components in each stage of our pipeline. In Stage 1, we found that leveraging contextual information from radiology reports (ConTEXTualNet) significantly improves localization performance compared to simpler architecture like U-Net. Moreover, the importance of text guidance is underscored by the drop in performance when constant or no text is provided.

In Stage 2, we investigated the effects of different configurations of the DAFT and varying grid resolutions. Our findings suggest that placing DAFT in the decoder layers is more effective, and a grid resolution of 4 achieves a good balance between detail capture and computational efficiency. Furthermore, the learned localization vector from Stage 1 is shown to be more effective than directly using segmentation masks.

Overall, our two-stage approach demonstrates superior segmentation performance, particularly in accurately localizing and segmenting the pneumothorax. By effectively integrating image and text information, our method provides a promising solution for improving pneumothorax diagnosis in clinical settings. Future work could explore further refinements and extensions of our approach, such as incorporating additional modalities or refining the attention mechanisms.

Chapter 4

Abdominal Multi-Organ Segmentation with Guided-nnUNet

4.1 Introduction

Multi-organ segmentation, a fundamental task in medical image analysis, involves delineating various organs simultaneously from imaging modalities such as CT and MRI. This process plays a crucial role in computer-aided diagnosis, treatment planning, and disease monitoring. However, multi-organ segmentation presents unique challenges not commonly encountered in general segmentation tasks. These challenges include navigating intricate spatial relationships between organs, where one organ may partially obscure the other, resulting in ambiguous boundaries. The varying shapes and sizes of organs within the same image, along with their proximity, also make it difficult to separate them accurately.

These challenges necessitate the use of automated and semi-automated segmentation approaches. DL techniques play a crucial role in multi-organ segmentation by harnessing the capabilities of neural networks to automatically delineate organs from medical images. Simple DL architectures, such as U-Net [6, 16] and V-Net [30] have been widely adopted for multi-organ segmentation because of their effectiveness in feature extraction. However, these methods often face challenges such as imbalanced classes and difficulty in distinguishing boundaries between adjacent organs, leading to under or over-segmentation. Moreover, the increased number of input and output channels necessary to represent various organs in the segmentation of 3D medical images aggravates the computational load and memory constraints, thereby affecting the efficacy and scalability of the algorithms. In addition, these models exhibit size bias, favoring larger organs in segmentation [31].

Patch-based training approaches such as nnUNet [8] address the issue of handling a large number of channels in medical images but still face limitations in handling complex anatomical structures and accurate segmentation, especially for organs in close proximity. More recent approaches like Label Conditioned Segmentation (LCS), address the challenge of segmenting images with a very large number of classes [32]. It achieves segmentation through a single-channel output regardless of the number of classes. LCS introduces an additional input, the *conditioned label* which is appended to the bottleneck layer of the model. During inference, this conditioned label acts as a guide and specifies the class of

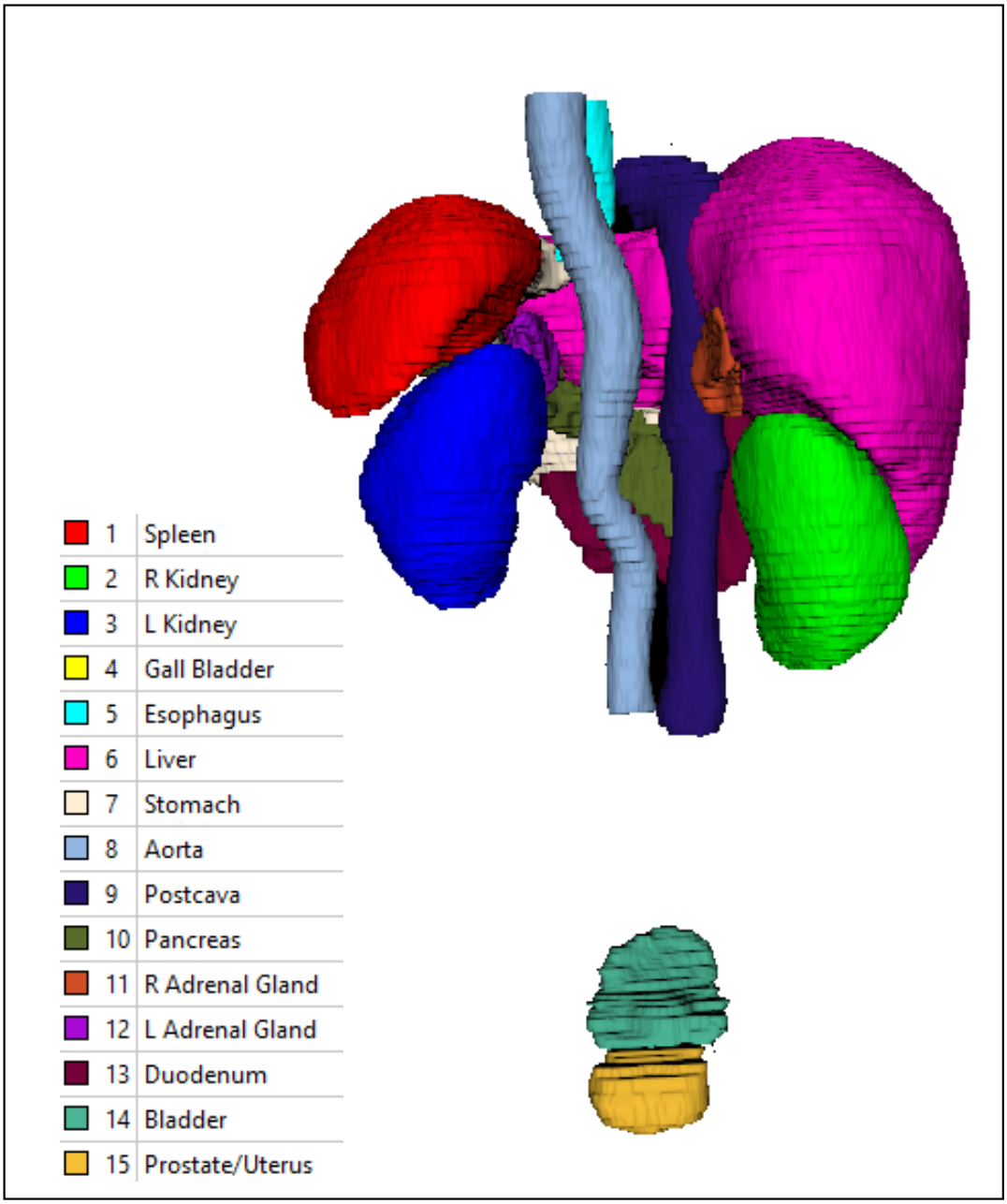


Figure 4.1 This illustration depicts the 15 abdominal organs included in the AMOS22 dataset for segmentation tasks.

interest for the segmentation. However, LCS is limited by its reliance on an explicit atlas designed for a set of classes, and segmenting multiple organs requires running inferences for each organ individually,

leading to significant computational overhead.

To overcome the challenges particular to anatomy, alternative approaches to explicit conditional modelling have been explored to improve segmentation accuracy. Topological Interaction Module proposed in [33] focuses on capturing spatial relationships like enclosing (one class being inside another) and exclusion (certain classes never appearing together) between organs. While effective for certain relationships, it may not cover some organ interactions like relative positioning, leading to potential segmentation inaccuracies and size bias. Conditional shape-location priors and unsupervised intensity priors [34] have been used to address shape, location and intensity variations. Despite the potential, it uses atlases and can struggle with complex anatomical structures or images with poor quality due to limitations in their ability to adapt to unseen variations.

In multi-organ segmentation, understanding the global context and relationships between all organs, not just immediate neighbors, is crucial. Therefore, transformers, known for their ability to capture long-range dependencies in data, are gaining attraction in medical image segmentation tasks [35–37]. However, their effectiveness is limited by the typically smaller medical image datasets compared to other domains. MedNeXt [38] proposes a bridge between DL and transformers by leveraging the strengths of both architectures, achieving state-of-the-art performance in multi-organ segmentation benchmarks. It utilizes ConvNeXt [39], a novel CNN architecture inspired by transformers. ConvNeXt incorporates transformer modules to capture long-range dependencies while retaining the efficient feature extraction capabilities of traditional CNNs.

While DL offers promise in multi-organ segmentation, existing single-stage models face inherent limitations. Unlike isolated objects, organs possess well-defined spatial relationships – the liver always sits next to the right kidney, for example. Single-stage models often overlook these crucial spatial cues, requiring complex algorithms to reason about the 3D layout and integrate this knowledge into the segmentation process. Furthermore, these models attempt to perform both coarse (identifying general organ areas) and fine-grained segmentation (accurately delineating boundaries) simultaneously, which can result in an information bottleneck leading to inaccurate boundaries, particularly when dealing with complex structures. In addition, single-stage frameworks may lack sufficient computational capacity for multiple complex anatomical structures and struggle to address class imbalances. Figure 4.1 shows the organs considered for this study.

Our proposed two-stage approach - Guided-nnUNet, overcomes these limitations by decomposing segmentation into two distinct stages. This allows the model to explicitly focus on the efficient localization of organs in the first stage and leverage this information for precise boundary delineation in the second stage. This division of tasks alleviates the information bottleneck and allows the model to better handle complex anatomical structures, ultimately leading to more accurate multi-organ segmen-

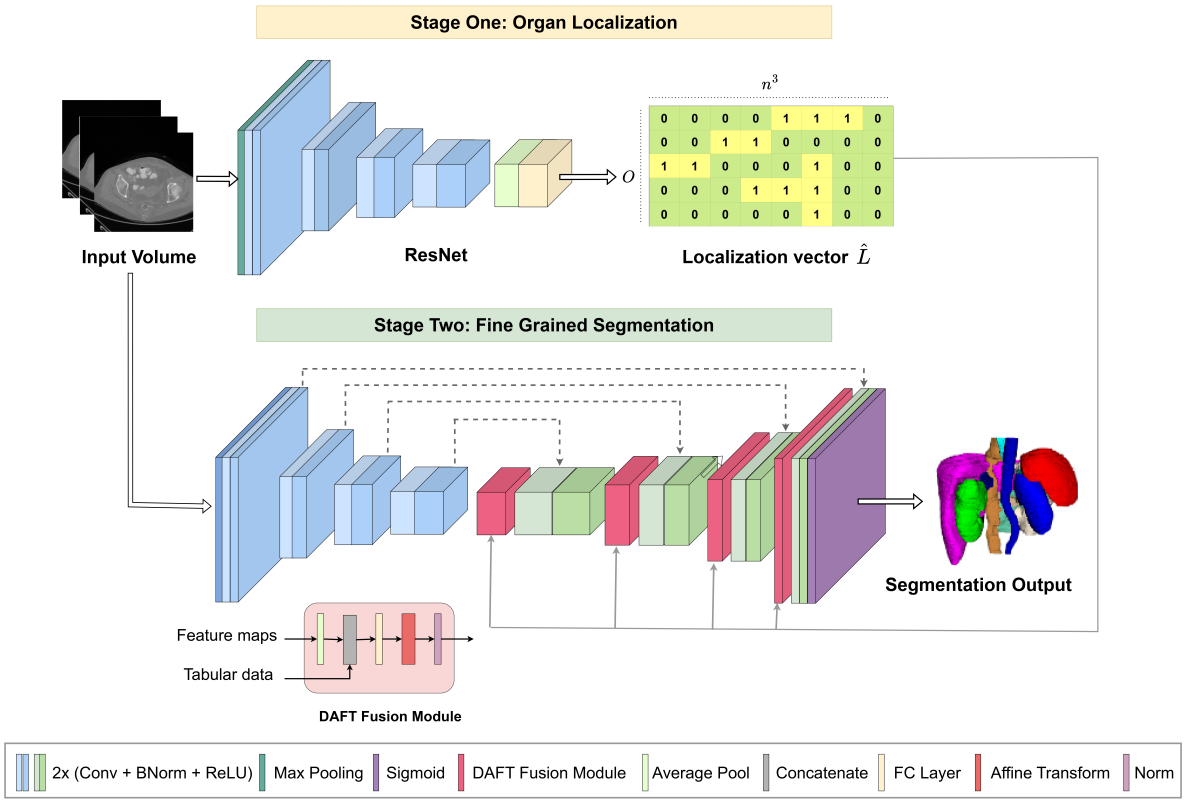


Figure 4.2 The proposed approach is illustrated in the schematic diagram. Initially, a coarse region localization map $\hat{\mathbf{L}}$ is predicted from the abdomen CT volume. Next, the feature maps of the segmentation network are modulated at various scales using $\hat{\mathbf{L}}$ to precisely segment the organs in the abdominal cavity.

tation. In the first stage, the model analyzes the input image and generates a low-dimensional map. This map highlights the approximate locations of the objects of interest within the image. This stage is computationally lighter as it deals with a simpler task. Next, in the second stage, we incorporate prior region maps at multiple scales within the encoder-decoder segmentation framework through DAFT [18]. DAFT predicts the scales and shifts to excite and suppress image feature maps of a convolutional layer by conditioning them on both the image and the map. To summarize, the key contributions of our work are:

- A novel, two-stage 3D segmentation framework, Guided-nnUNet, for accurate multi-organ segmentation in abdominal CT scans
- Using DAFT to fuse the organ localization information as spatial guidance to improve fine-grained organ segmentation

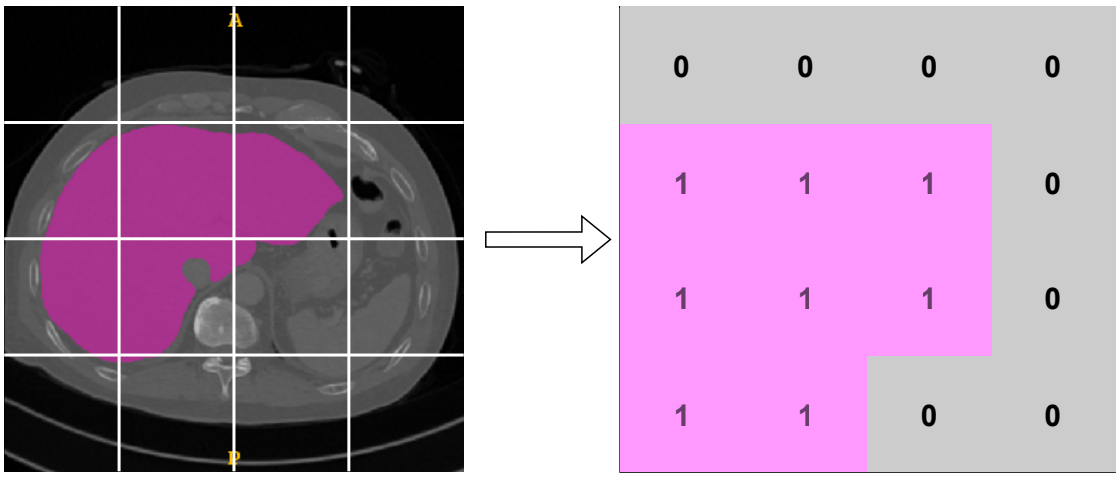


Figure 4.3 This figure shows a 2D illustration of a 3D localization map (n=4) generated in stage 1 for an axial sample slice for the liver. Similar localization maps are obtained, flattened and stacked organwise.

4.2 Proposed Method

The proposed Guided-nnUNet (shown in Figure 4.2) follows the pipeline introduced in Chapter 2, operating in two stages to achieve accurate and efficient organ segmentation in abdominal CT scans.

- **Stage 1: Organ Localization** - A ResNet-50 model [40] is employed to predict the presence or absence of each organ within individual image blocks (like a grid) across the entire scan. This initial step generates a low-dimensional localization map that functions as a spatial guide for the second stage, highlighting potential locations of each organ within the abdominal cavity.

- **Stage 2: Fine-grained Segmentation** - A 3D U-Net architecture equipped with DAFT is utilized in this stage. DAFT plays a vital role for effectively merging the rich visual features extracted from the CT volume with the broader spatial context provided by the localization map from stage 1. This combined information allows the network to focus its segmentation efforts on the most likely organ regions within the image, leading to improvements in segmentation performance.

4.2.1 Organ Localization

In the first stage of the framework, a simple classification architecture (ResNet [40]) is utilized to generate a preliminary map, guiding the subsequent segmentation stage. We feed the entire 3D image volume (abdominal CT scan) into the ResNet. The model then outputs a localization map $\hat{\mathbf{L}}$ with dimensions $O \times n^3$. The image volume is segmented into a 3D grid with a total of $n \times n \times n$ smaller

blocks. The localization map serves as a preliminary guide, highlighting potential regions where each organ might be located during the subsequent segmentation stage. Figure 4.3 shows a 2D illustration of the localization map generated in stage 1 for an axial sample slice for the liver.

Since this stage is formulated as a classification problem (predicting whether each block contains a specific organ or not), the network is trained using a cross-entropy loss function. Minimizing the cross-entropy loss function during training, trains the model to learn feature representations that effectively distinguish organ-containing regions from background areas and other organs within the CT volume. By learning these informative features, the ResNet model can generate a localization map that provides spatial context for the 3D U-Net architecture in the subsequent segmentation stage.

4.2.2 Fine-Grained Segmentation

This stage accurately segments the individual organs within the abdomen region. To achieve this, it leverages both the CT volume and the grid-level localization information obtained from the previous stage. The core architecture employed here is a 3D U-Net (specifically, the nnUNet implementation). nnUNet [8] is a well-established framework specifically designed for medical image segmentation tasks. It offers the benefits of a U-Net for accurate segmentation while leveraging pre-processing steps, a well-defined training pipeline, and pre-tuned hyperparameters, all provided by the framework. To further enhance segmentation performance, we incorporate DAFT modules [18] within the segmentation framework, as detailed in Section 2.4.2. DAFT guides the decoder towards a more accurate segmentation.

This stage utilizes a combination loss function. The loss function leverages the strengths of both Dice loss and Cross-Entropy loss. It addresses class imbalance concerns with Cross Entropy (CE) loss while promoting accurate segmentation boundary prediction through Dice loss, ultimately leading to a robust segmentation model for abdominal CT scans.

4.3 Experiments and Results

4.3.1 Dataset and Experimental Settings

4.3.1.1 Dataset

We evaluated all experiments using two publicly available datasets for abdominal multi-organ segmentation: AMOS22 [41] and BTCV. AMOS22 presents a rich collection of 500 CT scans and 100 MRI scans, sourced from various hospitals, scanner models, imaging modalities, and disease conditions. Each scan is meticulously segmented at the voxel level, covering a comprehensive list of 15 abdominal organs. Our focus for this study was solely on the CT scans from the AMOS22 dataset. The BTCV dataset, on the other hand, is specifically tailored for CT-based segmentation of abdominal organs. While smaller in size with only 30 CT volumes, BTCV offers meticulous segmentation for 13

Organs	nnUNet	Guided-nnUNet	MedNeXt	Guided-MedNeXt
Spleen	0.917 ± 0.019	$\mathbf{0.965 \pm 0.003}$	0.954 ± 0.008	0.961 ± 0.010
R Kidney	0.926 ± 0.015	$\mathbf{0.960 \pm 0.005}$	0.948 ± 0.003	0.959 ± 0.006
L Kidney	0.937 ± 0.016	$\mathbf{0.964 \pm 0.002}$	0.955 ± 0.005	0.963 ± 0.005
Gall Bladder	0.735 ± 0.027	0.826 ± 0.058	0.775 ± 0.046	$\mathbf{0.876 \pm 0.041}$
Esophagus	0.723 ± 0.005	$\mathbf{0.848 \pm 0.011}$	0.792 ± 0.026	0.804 ± 0.078
Liver	0.961 ± 0.007	0.971 ± 0.003	0.963 ± 0.004	$\mathbf{0.974 \pm 0.005}$
Stomach	0.850 ± 0.018	0.882 ± 0.013	0.866 ± 0.009	$\mathbf{0.903 \pm 0.027}$
Aorta	0.925 ± 0.008	$\mathbf{0.946 \pm 0.005}$	0.932 ± 0.005	0.937 ± 0.025
Postcava	0.846 ± 0.017	$\mathbf{0.909 \pm 0.002}$	0.882 ± 0.009	0.899 ± 0.019
Pancreas	$0.914 \pm 0.0.15$	$\mathbf{0.923 \pm 0.024}$	0.833 ± 0.014	0.876 ± 0.008
R Adrenal Gland	0.729 ± 0.036	0.724 ± 0.010	0.720 ± 0.052	$\mathbf{0.738 \pm 0.016}$
L Adrenal Gland	0.655 ± 0.037	0.762 ± 0.013	0.738 ± 0.028	$\mathbf{0.782 \pm 0.026}$
Duodenum	0.649 ± 0.020	0.825 ± 0.009	0.788 ± 0.021	$\mathbf{0.827 \pm 0.026}$
Bladder	0.832 ± 0.024	0.867 ± 0.057	0.809 ± 0.031	$\mathbf{0.893 \pm 0.019}$
Prostate/Uterus	0.776 ± 0.014	$\mathbf{0.855 \pm 0.026}$	0.809 ± 0.042	$\mathbf{0.855 \pm 0.020}$
Average	0.825 ± 0.006	0.882 ± 0.009	0.851 ± 0.010	$\mathbf{0.883 \pm 0.003}$

Table 4.1 Comparison of segmentation performance with (Guided-nnUNet, Guided-MedNeXt) and without (nnUNet, MedNeXt) localization guidance on the AMOS dataset. Five-fold average Dice scores along with standard deviation is reported.

abdominal organs. In our experiments, we employed a 5-fold cross-validation strategy on both datasets instead of utilizing the leaderboards.

4.3.1.2 Implementation

Our experiments utilized a two-stage approach. For stage 1, we resized the image volumes to a standard size of $64 \times 64 \times 64$ and employed a random five-fold cross-validation strategy. Within each fold, a designated test set was chosen, while the remaining data was further split into a 75% training set and a 25% validation set. The training process for stage 1 leveraged the AdamW optimizer with an initial learning rate and weight decay of 1e-4 each, implemented within the PyTorch framework. For stage 2, we transitioned to the default nnUNet pipeline with the 3d_fullres configuration. We reused the train-val-test split used in stage 1. The two stages were trained sequentially, with a maximum of 100 epochs for stage 1 and 1000 epochs for stage 2. Both stages utilized a batch size of 2 and were conducted on a system equipped with two NVIDIA GeForce RTX-2080Ti GPUs.

4.3.2 Results and Discussion

This study compares the segmentation performance of our two-stage pipeline with a baseline U-Net (nnUNet implementation) on two datasets: AMOS and BTCV.

Organs	nnUNet	Guided-nnUNet	MedNeXt	Guided-MedNeXt
Spleen	0.900 ± 0.046	$\mathbf{0.964 \pm 0.021}$	0.913 ± 0.040	0.917 ± 0.044
R Kidney	0.917 ± 0.019	$\mathbf{0.961 \pm 0.023}$	0.913 ± 0.043	0.929 ± 0.023
L Kidney	0.746 ± 0.020	$\mathbf{0.953 \pm 0.031}$	0.910 ± 0.040	0.915 ± 0.043
Gall Bladder	0.718 ± 0.067	$\mathbf{0.770 \pm 0.021}$	0.738 ± 0.022	0.767 ± 0.024
Esophagus	0.823 ± 0.021	$\mathbf{0.858 \pm 0.028}$	0.807 ± 0.018	0.825 ± 0.014
Liver	0.907 ± 0.005	$\mathbf{0.970 \pm 0.019}$	0.945 ± 0.019	0.950 ± 0.018
Stomach	0.882 ± 0.057	$\mathbf{0.947 \pm 0.016}$	0.879 ± 0.032	0.911 ± 0.016
Aorta	0.846 ± 0.017	$\mathbf{0.936 \pm 0.019}$	0.920 ± 0.009	0.929 ± 0.003
Postcava	0.852 ± 0.023	$\mathbf{0.896 \pm 0.002}$	0.869 ± 0.014	0.876 ± 0.016
Veins	0.732 ± 0.007	$\mathbf{0.835 \pm 0.044}$	0.753 ± 0.004	0.768 ± 0.041
Pancreas	0.750 ± 0.015	$\mathbf{0.865 \pm 0.032}$	0.825 ± 0.018	0.840 ± 0.015
R Adrenal Gland	0.757 ± 0.038	$\mathbf{0.796 \pm 0.006}$	0.747 ± 0.003	0.760 ± 0.028
L Adrenal Gland	0.767 ± 0.018	$\mathbf{0.815 \pm 0.034}$	0.754 ± 0.034	0.760 ± 0.034
Average	0.815 ± 0.018	$\mathbf{0.889 \pm 0.015}$	0.844 ± 0.023	0.857 ± 0.019

Table 4.2 Comparision of segmentation performance with (Guided-nnUNet, Guided-MedNeXt) and without (nnUNet, MedNeXt) localization guidance on the BTCV dataset. Five-fold average Dice scores along with standard deviation is reported.

Fusing DAFT with nnUNet resulted in significant improvements in organ segmentation performance across the two datasets as shown in Tables 4.1 and 4.2. The relative average Dice score increase was 7% for AMOS and 9% for BTCV. However, the impact of DAFT varied by organ size and complexity. In both datasets, the improvement was generally more pronounced for smaller organs with intricate features. For example, in AMOS, the increase ranged from a minimal 0.9% for the liver (a large and well-defined organ) to a maximum of 27% for the duodenum (a smaller and more challenging structure). This suggests that DAFT's spatial guidance from the localization map is particularly helpful for segmenting smaller organs where precise delineation is crucial. Although a similar pattern did not emerge in BTCV, with the improvement ranging from 4% for the esophagus to 27% for the left kidney (large organ), positive results were still observed.

This variation can be attributed to the information available in the CT scan and organ size. For larger organs with distinct features, nnUNet might perform well on its own. However, for smaller organs with limited visual cues in the CT scan, DAFT's additional spatial guidance becomes essential, leading to more significant accuracy gains. The larger and more diverse AMOS dataset provides a more comprehensive picture of this effect compared to the potentially limited size of the BTCV dataset.

In Table 4.3, we see that the improvement was most significant for the duodenum (27.096%), esophagus (17.294%), and left adrenal gland (16.265%) in AMOS, which are all smaller organs with less prominent features in CT scans. Organs like the liver (0.978%) and aorta (2.291%) showed a smaller improvement due to their larger size and distinct features. An interesting exception to this trend is the

Organs	% of Voxels	% Improvement over nnUNet
R Adrenal Gland	0.045	−0.714
L Adrenal Gland	0.050	16.265
Esophagus	0.207	17.294
Gall Bladder	0.416	12.415
Prostate/Uterus	0.642	10.178
Duodenum	0.782	27.096
Postcava	0.935	7.450
Pancreas	1.05	1.007
Bladder	1.537	4.233
Aorta	1.61	2.291
Liver	1.87	0.978
R Kidney	2.110	3.649
L Kidney	2.185	2.883
Spleen	2.110	3.649
Stomach	4.58	3.765
Average	-	6.911

Table 4.3 Percentage of average voxels occupied by each organ in the AMOS dataset and the average Dice score improvement achieved by Guided-nnUNet compared to nnUNet for each organ.

right adrenal gland. Medical literature confirms that the right adrenal gland is inherently smaller and more challenging to identify compared to the left one, due to its anatomical position behind a large blood vessel and potential obscuration by the liver [42, 43]. As a result, the model's performance with or without DAFT shows minimal variation for this specific gland.

4.3.2.1 Comparison with state-of-the-art method

As shown in Table 4.1 and Table 4.2, our analysis of Guided-nnUNet and the SOTA MedNeXt for abdominal organ segmentation reveals a clear advantage for Guided-nnUNet. On both AMOS and BTCV datasets, Guided-nnUNet achieves a higher average Dice score compared to MedNeXt. This improvement ranges from 3.6% on the AMOS dataset to 5.3% on the BTCV dataset, and is particularly notable for smaller organs, such as the pancreas (10.81%), esophagus (7.07%), and gallbladder (6.58%). Even for organs like the liver and kidneys, there is a clear advantage for Guided-nnUNet. It is worth noting that the right adrenal gland, a small and inherently difficult organ to segment, shows the least improvement (0.56%).

Building on the success of MedNeXt, a SOTA method for abdominal organ segmentation, we investigated the impact of incorporating DAFT. Adding DAFT into MedNeXt (referred to as Guided-MedNeXt) demonstrates improvement in segmentation performance across various organs. Overall, it improves upon MedNeXt on AMOS by 3.7% and on BTCV by 1.5%. Notably, the improvement is more

Organs	nnUNet	Guided-nnUNet (fixed loc)	Guided-nnUNet
Spleen	0.900 ± 0.046	0.912 ± 0.002	$\mathbf{0.964 \pm 0.021}$
R Kidney	0.917 ± 0.019	0.898 ± 0.010	$\mathbf{0.961 \pm 0.023}$
L Kidney	0.746 ± 0.020	0.737 ± 0.034	$\mathbf{0.953 \pm 0.031}$
Gall Bladder	0.718 ± 0.067	0.727 ± 0.016	$\mathbf{0.770 \pm 0.021}$
Esophagus	0.823 ± 0.021	0.819 ± 0.008	$\mathbf{0.858 \pm 0.028}$
Liver	0.907 ± 0.005	0.920 ± 0.012	$\mathbf{0.970 \pm 0.019}$
Stomach	0.882 ± 0.057	0.897 ± 0.029	$\mathbf{0.947 \pm 0.016}$
Aorta	0.846 ± 0.017	0.828 ± 0.010	$\mathbf{0.936 \pm 0.019}$
Postcava	0.852 ± 0.023	0.833 ± 0.019	$\mathbf{0.896 \pm 0.002}$
Veins	0.732 ± 0.007	0.741 ± 0.013	$\mathbf{0.835 \pm 0.044}$
Pancreas	0.750 ± 0.015	0.750 ± 0.027	$\mathbf{0.865 \pm 0.032}$
R Adrenal Gland	0.757 ± 0.038	0.744 ± 0.027	$\mathbf{0.796 \pm 0.006}$
L Adrenal Gland	0.767 ± 0.018	0.760 ± 0.022	$\mathbf{0.815 \pm 0.034}$
Average	0.815 ± 0.018	0.813 ± 0.002	$\mathbf{0.889 \pm 0.015}$

Table 4.4 Comparison of segmentation performance (Dice similarity coefficient) on the BTCV dataset, showing nnUNet, Guided-nnUNet (fixed loc) with fixed localization, and Guided-nnUNet with dynamic localization.

pronounced for smaller organs like the pancreas, gallbladder, duodenum, and left adrenal gland, ranging from 4.95% to 13%.

4.3.3 Ablation Studies

4.3.3.1 Impact of Localization

In the standard Guided-nnUNet approach, stage 1 generates a localization map that provides spatial information about the target organs. However, in this experiment, a constant vector with all 1s was fed into stage 2 instead of the stage 1 output. This essentially bypasses the stage 1 localization step and provides no specific guidance about organ location. Table 4.4 compares the performance of three models:

- nnUNet: Baseline model without DAFT.
- Guided-nnUNet (fixed loc): This version uses Guided-nnUNet with the constant localization vector.
- Guided-nnUNet: This refers to the standard Guided-nnUNet model with the localization map generated by stage 1.

The results in Table 4.4 show a mixed effect of using the constant localization vector. For organs like the spleen (0.912), liver (0.920), and stomach (0.897), the Guided-nnUNet (fixed loc) achieved a

Organs	nnUNet	Guided-nnUNet (n=2)	Guided-nnUNet (n=3)	Guided-nnUNet (n=4)
Spleen	0.900 ± 0.046	0.932 ± 0.008	0.943 ± 0.008	$\mathbf{0.964 \pm 0.021}$
R Kidney	0.917 ± 0.019	0.932 ± 0.016	0.942 ± 0.016	$\mathbf{0.961 \pm 0.023}$
L Kidney	0.746 ± 0.020	0.893 ± 0.043	0.937 ± 0.043	$\mathbf{0.953 \pm 0.031}$
Gall Bladder	0.718 ± 0.067	0.735 ± 0.025	0.761 ± 0.025	$\mathbf{0.770 \pm 0.021}$
Esophagus	0.823 ± 0.021	0.856 ± 0.012	$\mathbf{0.858 \pm 0.012}$	$\mathbf{0.858 \pm 0.028}$
Liver	0.907 ± 0.005	0.944 ± 0.022	0.959 ± 0.022	$\mathbf{0.970 \pm 0.019}$
Stomach	0.882 ± 0.057	0.934 ± 0.003	0.937 ± 0.003	$\mathbf{0.947 \pm 0.016}$
Aorta	0.846 ± 0.017	0.852 ± 0.008	0.912 ± 0.008	$\mathbf{0.936 \pm 0.019}$
Postcava	0.852 ± 0.023	0.857 ± 0.008	0.882 ± 0.008	$\mathbf{0.896 \pm 0.002}$
Veins	0.732 ± 0.007	0.739 ± 0.036	0.818 ± 0.019	$\mathbf{0.835 \pm 0.044}$
Pancreas	0.750 ± 0.015	0.754 ± 0.036	0.845 ± 0.036	$\mathbf{0.865 \pm 0.032}$
R Adrenal Gland	0.757 ± 0.038	0.750 ± 0.024	0.790 ± 0.024	$\mathbf{0.796 \pm 0.006}$
L Adrenal Gland	0.767 ± 0.018	0.770 ± 0.044	0.788 ± 0.044	$\mathbf{0.815 \pm 0.034}$
Average	0.815 ± 0.018	0.842 ± 0.013	0.875 ± 0.022	$\mathbf{0.889 \pm 0.015}$

Table 4.5 Comparing segmentation performance (Dice similarity coefficient) on BTCV dataset using Guided-nnUNet with different grid block configurations (n=2, 3, 4)

slight improvement compared to the baseline nnUNet. This suggests that even without specific localization information, DAFT might still provide some benefit by incorporating additional features from the localization map. Whereas for organs like the kidneys (right and left), esophagus, aorta, and adrenal glands, the constant localization vector resulted in either similar or slightly worse performance compared to nnUNet alone. This indicates that the spatial guidance from the localization map in the standard Guided-nnUNet is crucial for accurate segmentation for these organs.

The average Dice score for Guided-nnUNet (fixed loc) (0.813) is lower by 8.55% compared to the standard Guided-nnUNet (0.889). This highlights the importance of the stage 1 localization step in the overall performance of the model, especially for organs where precise delineation is necessary.

4.3.3.2 Impact of Grid Resolution

Table 4.5 explores how the parameter n affects the performance of our segmentation model (Guided-nnUNet) on the BTCV dataset. As discussed in Section 4.2.2, n controls the number of squares (grid blocks) used to divide each image. A higher number of grid blocks (larger n) creates a finer grid. The results show that increasing n from 2 to 3 generally leads to better segmentation accuracy, as measured by Dice score, for most organs. This suggests that a finer grid allows the model to capture the details of these organs more effectively.

However, the impact of n is not uniform across all organs. Some organs like the esophagus exhibit no significant change in performance even with a finer grid (n=4). This suggests that for these organs,

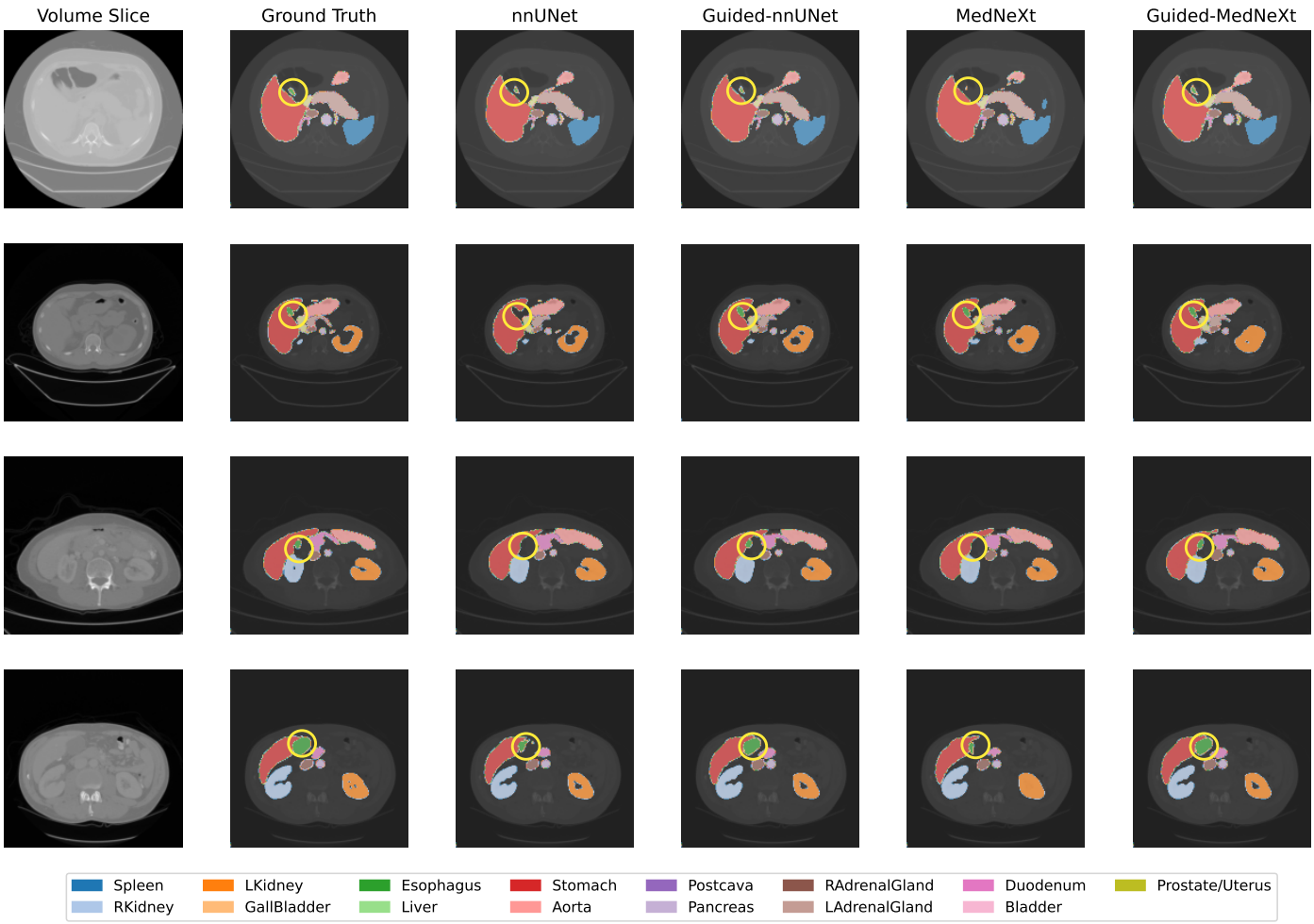


Figure 4.4 The performance of various segmentation algorithms on abdominal organs in CT scans is compared qualitatively. Specifically, the liver and gallbladder are the focus of this analysis. The image slice, ground truth, nnUNet, Guided-nnUNet, MedNeXt, and Guided-MedNeXt results are displayed in each row.

a finer grid might not be advantageous. There are two possible explanations: (i) Diminishing return: excessively small grid blocks might become less informative for capturing the relevant features of these organs, (ii) Overfitting: with a higher number of grid blocks (n=4), the model might be focusing on irrelevant details in the training data, leading to a decrease in generalizability on unseen test data.

Considering these factors, a range of n=3 to 4 appears to be ideal. This range allows the model to capture detailed features while keeping the number of parameters in check and reducing the risk of overfitting. However, optimal n can vary depending on the specific dataset and organs being analyzed.

4.3.3.3 Qualitative Analysis

Figures 4.4 and 4.5 offer some insights into the segmentation performance of different approaches for challenging abdominal organ structures. For this analysis, two specific pairs of organs are highlighted: (i) liver and gall bladder and (ii) stomach and spleen. These closely located organs often exhibit intricate boundaries and similar intensities in CT scans, making segmentation difficult. As seen in Figure 4.4 and Figure 4.5, visual inspection reveals that nnUNet struggles to differentiate between them, potentially leading to under-segmentation of the gallbladder. Guided-nnUNet, on the other hand, appears to achieve clearer separation due to the guidance provided by the localization map. Similarly, MedNeXt might exhibit challenges in accurately delineating the gallbladder, while Guided-MedNeXt could offer a more precise segmentation.

4.4 Conclusion

This work investigated an application of the two-stage segmentation framework introduced in Chapter 2 - Guided-nnUNet, for abdominal organs in CT scans. Guided-nnUNet decomposes the segmentation process into two steps: organ localization and then localization-guided fine-grained segmentation. Guided-nnUNet incorporates a spatial guidance module (DAFT) that significantly improves segmentation accuracy, especially for small and intricate organs, compared to the baseline nnUNet. The benefit of DAFT was more pronounced in the larger and more diverse AMOS dataset. Ablation studies confirmed the importance of DAFT's guidance and grid resolution for precise segmentation. Notably, Guided-nnUNet outperformed MedNeXt on both datasets, highlighting its effectiveness. While this study focused on Guided-nnUNet, combining DAFT with other models like MedNeXt shows scope for further improvements, requiring future exploration.

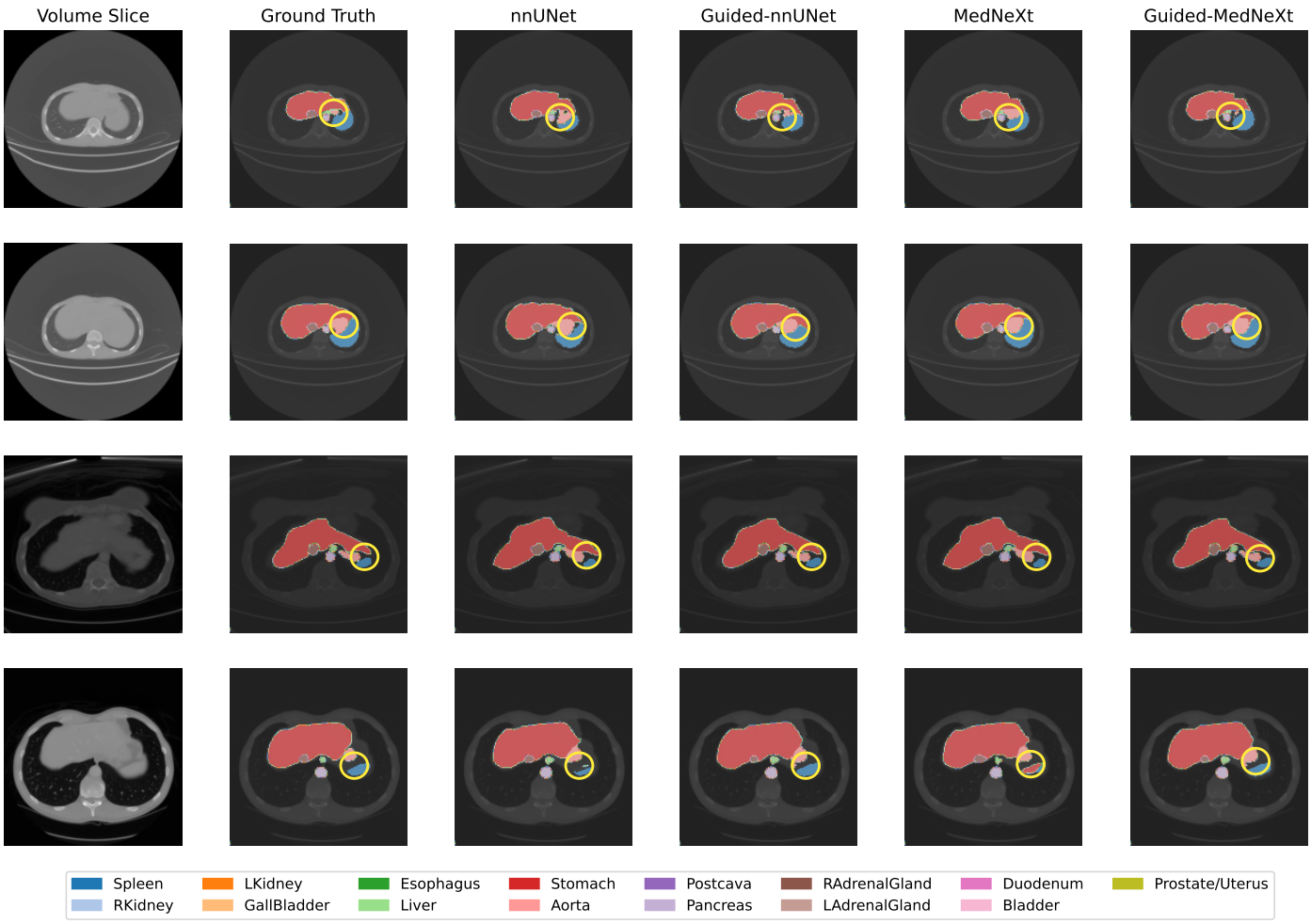


Figure 4.5 The performance of various segmentation algorithms on abdominal organs in CT scans is compared qualitatively. Specifically, the stomach and spleen are the focus of this analysis. The image slice, ground truth, nnUNet, Guided-nnUNet, MedNeXt, and Guided-MedNeXt results are displayed in each row.

Chapter 5

Conclusion

This thesis proposes a novel two-stage DL pipeline for medical image segmentation, addressing the challenges posed by single-stage models. The thesis demonstrates the efficacy of the pipeline in improving segmentation performance by separating the tasks of localization and segmentation. In the first stage, the pipeline focuses on localizing target organs/lesions. The second stage utilizes DAFT-enhancement to perform fine-grained segmentation, focusing on the most relevant image areas. This method results in a more robust and efficient approach to medical image segmentation, particularly for complex tasks involving anatomical variations and overlapping organs.

The effectiveness of this two-stage approach has been demonstrated through two applications. One application involved accurate pneumothorax segmentation in chest radiographs. It was shown to outperform existing methods like U-Net, LViT, ConTEXTualNet and CPAM through a novel approach that leverages additional information sources, such as free-text radiology reports, which was introduced to enhance the accuracy of pneumothorax segmentation. The ablation studies highlighted the importance of contextual information from radiology reports and the effectiveness of DAFT modules in the decoder layers.

Another application of the pipeline for abdominal organ segmentation in CT scans was demonstrated. Guided-nnUNet, which incorporates DAFT modules, achieved superior segmentation accuracy compared to the baseline nnUNet, particularly for small and intricate organs. This further emphasizes the benefits of the two-stage approach and DAFT modules in improving segmentation performance. Ablation studies confirm the importance of DAFT's guidance and the choice of grid resolution in achieving precise segmentation. The findings suggest that the integration of DAFT with other models could lead to further enhancements in segmentation accuracy.

The versatility of the proposed two-stage pipeline is further highlighted by its successful application in tasks dealing with both 2D and 3D medical images. One application focused on accurate pneumothorax segmentation in chest radiographs, which are inherently 2D images. In contrast, the other application focused on abdominal organ segmentation in CT scans, representing 3D volumes. This adaptability

showcases the pipeline's ability to achieve accurate segmentation regardless of image dimensionality, further strengthening its potential for broader applications in diverse medical image analysis tasks.

5.1 Limitations

While the two-stage pipeline offers a robust approach to medical image segmentation, inherent limitations exist within each stage. Its increased complexity compared to single-stage models, requires careful optimization for both stages to ensure efficient and accurate segmentation. The pipeline incurs a higher computational cost due to running two separate stages, which could be a concern for real-time applications or resource-constrained environments. Grid-based localization in Stage 1 can be computationally expensive with high resolutions and susceptible to overfitting with very fine grids. Furthermore, the effectiveness of the pipeline relies on the quality and availability of training data, as insufficient data or data with significant noise might hinder the performance of both localization and segmentation stages.

5.2 Future work

Overall, the research contributes to the field of medical image analysis by providing a scalable and effective solution for complex segmentation tasks. Building upon the pipeline's success, future work can delve into several directions to enhance its capabilities and broaden its applications. Explainable AI (XAI) techniques could be integrated to understand how the model leverages information and makes predictions. Alternative localization methods like heatmap regression or keypoint detection hold promise for richer localization information. The framework's generalizability to diverse tasks like multi-task learning scenarios can be investigated. Furthermore, seamless clinical integration through user interfaces and electronic health record incorporation could be explored. Finally, leveraging unsupervised or weakly supervised learning, incorporating additional modalities like clinical notes or other imaging techniques, could further enhance segmentation performance and understanding. The two-stage pipeline holds promise for future applications in medical diagnostics, treatment planning, and disease monitoring, potentially revolutionizing the way medical images are interpreted and utilized in healthcare.

List of Related Publications

[P1] S. Shastri, N Akash RJ, L. Gautham and J. Sivaswamy **"Locate-then-Delineate: A Free-text Report Guided Approach for Pneumothorax Segmentation in Chest Radiographs"**, Accepted in *IEEE 21st International Symposium on Biomedical Imaging (ISBI)*, 2024.

[P2] S. Shastri, N Akash RJ, and J. Sivaswamy **"Leveraging Spatial Guidance for Accurate Multi-Organ Segmentation in Abdominal CT"**, submitted to *27th International Conference on Pattern Recognition (ICPR)*, 2024.

Bibliography

[1] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[2] R. Hemalatha, T. Thamizhvani, A. J. A. Dhivya, J. E. Joseph, B. Babu, and R. Chandrasekaran, "Active contour based segmentation techniques for medical image analysis," *Medical and Biological Image Analysis*, vol. 4, no. 17, p. 2, 2018.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[7] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Deepmedic for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2*. Springer, 2016, pp. 138–149.

[8] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[9] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[10] H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang, "Transformers in medical image segmentation: A review," *Biomedical Signal Processing and Control*, vol. 84, p. 104791, 2023.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[12] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] S. Feng, Q. Liu, A. Patel, S. U. Bazai, C.-K. Jin, J. S. Kim, M. Sarrafzadeh, D. Azzollini, J. Yeoh, E. Kim *et al.*, "Automated pneumothorax triaging in chest x-rays in the new zealand population using deep-learning algorithms," *Journal of medical imaging and radiation oncology*, vol. 66, no. 8, pp. 1035–1043, 2022.

[15] T. Dang, T. T. Nguyen, J. McCall, E. Elyan, and C. F. Moreno-García, "Two-layer ensemble of deep learning models for medical image segmentation," *Cognitive Computation*, pp. 1–20, 2024.

[16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[17] C. Liu, F. Denzinger, L. Folle, J. Qiu, L. Klein, J. Maier, M. Kachelrieβ, M. Lell, and A. Maier, "Whole-body multi-organ segmentation using anatomical attention," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[18] T. N. Wolf, S. Pölsterl, C. Wachinger, A. D. N. Initiative *et al.*, "Daft: A universal module to interweave tabular data and 3d images in cnns," *NeuroImage*, vol. 260, p. 119505, 2022.

[19] J. Tschopp, R. Rami-Porta, M. Noppen, and P. Astoul, "Management of spontaneous pneumothorax: state of the art," *European Respiratory Journal*, vol. 28, no. 3, pp. 637–650, 2006.

[20] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.

[21] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3942–3951.

[22] R. Bhalodia, A. Hatamizadeh, L. Tam, Z. Xu, X. Wang, E. Turkbey, and D. Xu, "Improving pneumonia localization via cross-attention on medical images and reports," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 571–581.

[23] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 155–18 165.

[24] Z. Li *et al.*, "Lvit: Language meets vision transformer in medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[25] G.-E. Lee *et al.*, "Text-guided cross-position attention for segmentation: Case of medical image," in *Medical Image Computing and Computer-Assisted Intervention*, 2023.

[26] Z. Huemann *et al.*, "Contextual net: A multimodal vision-language model for segmentation of pneumothorax," *arXiv preprint arXiv:2303.01615*, 2023.

[27] S. E. Mason *et al.*, "Semi-quantitative visual assessment of chest radiography is associated with clinical outcomes in critically ill patients," *Respiratory Research*, vol. 20, no. 1, pp. 1–9, 2019.

[28] B. A, P. A, K. E, I. V, and K. A, "Albumentations: fast and flexible image augmentations," *ArXiv e-prints*, 2018.

[29] A.-M. Kelly *et al.*, "Comparison between two methods for estimating pneumothorax size from chest x-rays," *Respiratory Medicine*, vol. 100, no. 8, pp. 1356–1359, 2006.

[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)*. Ieee, 2016, pp. 565–571.

[31] V. Mythri, A. J. Thottupattu, N. A. RJ, and J. Sivaswamy, "A method to remove size bias in sub-cortical structure segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–4.

[32] T. Ma, B. C. Lee, and M. R. Sabuncu, "Label conditioned segmentation," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 847–857.

[33] S. Gupta, X. Hu, J. Kaan, M. Jin, M. Mpoy, K. Chung, G. Singh, M. Saltz, T. Kurc, J. Saltz *et al.*, "Learning topological interactions for multi-class medical image segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 701–718.

[34] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from ct images using conditional shape–location and unsupervised intensity priors," *Medical image analysis*, vol. 26, no. 1, pp. 1–18, 2015.

[35] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[36] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[37] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, and F. S. Khan, "Unetr++: delving into efficient and accurate 3d medical image segmentation," *arXiv preprint arXiv:2212.04497*, 2022.

[38] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. H. Maier-Hein, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 405–415.

[39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 722–36 732, 2022.

[42] J. Vincent, I. Morrison, P. Armstrong, and R. Reznek, "The size of normal adrenal glands on computed tomography," *Clinical radiology*, vol. 49, no. 7, pp. 453–455, 1994.

[43] A. Harvey and E. Friend, "Adrenal gland," *Feline Soft Tissue and General Surgery*, pp. 393–399, 2014.