BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Multi-Organ and Tumor Segmentation by Context-Aware Attention with BioBERT and Global Spatial Reasoning

Mirza Mohammad Azwad
200042121
Rhidwan Rashid
200042149
M M Nazmul Hossain
200042118

Department of Computer Science and Engineering
Islamic University of Technology
April, 2025

Multi-Organ and Tumor Segmentation by Context-Aware Attention with BioBERT and Global Spatial Reasoning

Mirza Mohammad Azwad
200042121
Rhidwan Rashid
200042149
M M Nazmul Hossain
200042118

Department of Computer Science and Engineering
Islamic University of Technology
April, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Mirza Mohammad Azwad**, **Rhidwan Rashid**, and **M M Nazmul Hossain** under the supervision of **Tareque Mohmud Chowdhury**, Assistant Professor, Department of Computer Science and Engineering and co-supervision of **Farzana Tabassum**, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: April 27, 2025

Farzana Tabassum
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: April 27, 2025

Mirza Mohammad Azwad
Student ID: 200042121
Date: April 27, 2025

Rhidwan Rashid
Student ID: 200042149
Date: April 27, 2025

M M Nazmul Hossain
Student ID: 200042118
Date: April 27, 2025

Dedicated to our supervisor, whose guidance and support made this research journey smooth and successful.

Contents

List of Figures

List of Abbreviations

CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
nnU-net	Not a New U-net
DoDNet	Dynamic On Demand Network
FUSE	Fusion and Selection
BraTS	Brain Tumor Segmentation
MOTS	Multi Organ Tumor Segmentation

Acknowledgement

We would like to express our deepest gratitude to our supervisor, Mr. Tareque Mohmud Chowdhury, and our co-supervisor, Ms. Farzana Tabassum, Lecturer, Bioinformatics Lab, Department of Computer Science and Engineering, Islamic University of Technology, for their continuous guidance, valuable feedback, and unwavering support throughout the course of this project. Their expertise and encouragement have been instrumental in shaping our work.

We also wish to extend our sincere thanks to the Department of Computer Science and Engineering at Islamic University of Technology for providing us with the necessary resources and an inspiring academic environment.

Abstract

Accurate segmentation of multiple organs and tumors in medical images is critical for enhancing clinical workflows, yet existing methods often struggle with capturing long-range spatial dependencies, leveraging domain-specific semantics, and generalizing across diverse imaging modalities. This thesis addresses these challenges by proposing a novel hybrid architecture that integrates context-aware attention mechanisms with biomedical semantic embeddings to achieve robust multi-organ and tumor segmentation.

The proposed framework combines a CNN-Transformer hybrid encoder for local and global feature extraction, a *BioBERT*-enhanced universal prompt module to infuse domain-specific knowledge, and a *FUSE* prompt decoder for task-aware feature fusion. By incorporating *BioBERT*—a biomedical language model—the model enriches task prompts with anatomical and clinical semantics, enabling better understanding of organ-tumor relationships. The *TransAttUNet*-inspired encoder-decoder backbone, augmented with multi-scale skip connections, ensures precise reconstruction of segmentation masks while addressing spatial complexity.

Three diverse datasets: *MOTS* (abdominal CT), *BraTS 2021* (brain MRI), and *Prostate MRI*—are utilized to evaluate the model's performance across modalities and anatomical regions. These datasets present unique challenges, including interclass imbalance, multimodal data integration, and low-context scenarios, providing a comprehensive testbed for assessing generalization capabilities.

Chapter 1

Introduction

Medical image segmentation is the process of identifying and delineating anatomical structures or regions of interest (such as organs, tissues, or lesions) within biomedical images. In clinical diagnosis and planning, segmentation serves to "delineate the objects of interest from the complex background on various biomedical images"[1]. A special case of this is multi-organ segmentation, where multiple anatomical structures (and potentially associated pathological regions) are segmented simultaneously. For example, in cancer treatment planning one often needs both a tumor mass and its surrounding organs-at-risk (OARs) outlined in a scan[2]. Similarly, tumor detection refers to identifying suspicious regions that correspond to tumors; when combined with segmentation, it provides precise boundaries of tumors for analysis and intervention. These tasks are highly relevant to healthcare. Accurate organ and tumor segmentation can greatly aid computer-aided diagnosis, surgical planning, and radiotherapy, improving treatment outcomes and patient safety[2]. Thanks to advances in artificial intelligence and deep learning, fully automated segmentation methods have dramatically improved in recent years. In fact, modern deep learning–based segmentation models now "far outperforms traditional methods" and have become a major research topic[2]. This has paved the way for integrating AI tools into clinical workflows, helping to reduce manual workload and inter-observer variability.

1.1 Motivations and Scope

The motivation for this research arises from the need of accurate, efficient, and reliable segmentation of organs and tumors in medical images. For example, in radiation therapy for cancer, precise delineation of tumor volumes and nearby normal organs is essential to target tumor tissue while sparing healthy tissue.

Currently, segmentation is often performed manually by clinicians, which is time-consuming (e.g., contouring 24 organs in head-and-neck CT can take over 3 hours, labor-intensive, and subject to differences. Also, a shortage of trained experts can lead to delays in treatment and inconsistent results. Automated segmentation promises to ease this process, enabling faster treatment planning and more standardized outcomes.

Despite many progresses, existing segmentation models have limitations that motivate further research. Zhang et al. proposed a Dynamic On-demand Network (DoD-Net) that tackles partially labeled multi-organ data by using task-conditioned filters and a dynamic head[3]. On the other hand, Liu et al. introduced a CLIP-Driven Universal Model that uses text embeddings from a vision-language model (CLIP) to encode anatomical relationships, enabling segmentation of 25 organs and 6 tumor types[4]. However, because CLIP is pretrained on natural images, its medical generalization is limited.

Deep learning approaches like U-Net and its variants have achieved significant performance, but they usually focus on individual organs or single imaging modalities. For example, the nnU-Net framework (a self-configuring U-Net) has set high benchmarks across diverse medical segmentation challenges [5], but it relies only on convolutional operations and local context. Such CNN-based models may struggle to capture long-range spatial relationships, which can be important when organs or tumors span large/very small regions of the image. Recent transformer-augmented networks (such as TransAttUNet[1]) attempts to address this by adding attention modules, but these models can be computationally heavy and still may not fully leverage prior anatomical knowledge.

These observations reveal gaps in the current literature. Existing approaches tend to be specialized (focusing on one organ or modality) or require task labels that ignore rich anatomical relationships. There is no unified framework that both captures global image context and incorporates domain-specific knowledge for segmentation. In this thesis, we aim to address these gaps by developing methods that bridge CNN and transformer techniques, leverage multi-modal data, and integrate biomedical semantics into the segmentation process.

1.2 Problem Statement

Given a collection of N datasets $\{D_1, D_2, \ldots, D_N\}$, where each $D_i = \{(X_{ij}, Y_{ij})\}_{j=1}^{n_i}$ consists of n_i medical images X_{ij} and their corresponding ground truth segmentations Y_{ij}, the goal is to perform segmentation of multiple organs and tumors across diverse imaging modalities and anatomical regions.

1.3 Research Challenges

- **Difficulty in capturing long-range spatial dependencies using CNN-based backbones like nnU-Net:** Classical convolutional architectures like U-Net and nnU-Net have small receptive fields and are based heavily on local information. While this is fine for small or confined anatomical structures, it is rather challenging if dealing with spatially extensive or complex regions (such as tumors that cross the boundaries of organs). The former is suboptimal for capturing long-range structure using these models without explicit architectural modifications.
- **Inability to leverage biomedical domain knowledge effectively in segmentation prompts:** Recent approaches leveraging vision-language models or prompt guided embeddings have unlocked new possibilities, but there is a catch: most of them rely on general-purpose models such as CLIP, which were not trained to understand medical concepts. Therefore, the anatomical and clinical semantics are not effectively represented with prompts, and segmentation models may not understand organ-tumor interactions or typical spatial arrangements.
- **High domain shift sensitivity when transferring models across datasets or imaging modalities:** A model which was trained on abdominal CT might perform poorly on brain MRI because of domain shifts in intensity distributions, resolution, organ appearance, and annotation protocols. This prevents generalization, and requires either domain-specific fine-tuning or robust architectures.
- **Limited interpretability in prompt-guided segmentation due to opaque fusion mechanisms:** Though prompts are used to guide segmentation in prompt-driven segmentations, it is often unclear how the prompts influence the outcome. The fusion mechanism in many existing models are treated as a black box, which makes it difficult to debug, refine, and understand the decision-making process. This lack of interpretability becomes a concern in clinical set-

tings where transparency is crucial.

1.4 Contribution

This research proposes a novel hybrid segmentation pipeline that integrates a Transformer-based encoder (TransAttU-Net) [1]) with a BioBERT-driven prompt decoder.

The key contributions of this research are summarized as follows:

- **Development of a domain-aware universal segmentation framework** that bridges global spatial reasoning with biomedical semantic understanding.
- **Critical evaluation of state-of-the-art prompt-driven models:** Through a comparative analysis of existing universal segmentation models such as DoD-Net [3], CLIP-driven segmentation models [4], and UniSeg [6], several architectural shortcomings were identified. These include delayed prompt fusion, limited global spatial modeling, and inadequate semantic grounding using general-purpose embeddings.

1.5 Organization

The rest of the thesis is organized as follows:

- **Chapter 2: Related Works** In this chapter, we surveyed existing works on multi-organ tumor segmentation. It also includes traditional algorithms and deep learning methods. We mainly focused on recent architectures, such as U-Net variants, transformer based networks, prompt driven architectures etc. and discussed various approaches.
- **Chapter 3: Proposed Methodology** Here we presented our proposed pipeline. It includes out innovative pipeline for multi-organ tumor segmentation and describes it in detail.
- **Chapter 4: Results and Discussion** This chapter describes and and addresses the challenges of the datasets(MOTS, BraTS21, ProstateMRI) we are using for our thesis. It also includes insight in each dataset and their unique features.
- **Chapter 5: Conclusion** In this chapter, we summarize everything we have done until now restating the contributions and future directions of our thesis.

Chapter 2

Related Works

During the last decade, with the increased novelty of deep learning architectures and the advent of attention mechanisms [7], continuous efforts have been made to integrate deep learning architectures to create automated clinical workflows. Furthermore, advances in medical image segmentation have been pivotal for automating these clinical workflows with the goal of multi-organ and tumor detection. Since its inception, the research has evolved from task-specific models to universal frameworks capable of handling diverse organs and tumors across different image modalities, i.e., CT scans, MRI scans, PET scans, etc. In this section, we present a narrative that traces this evolution, highlighting the foundational architectures, dynamic models, prompt-driven segmentation, and their respective challenges and contributions. In Figure 2.1, we showcase the evolution of the field and our focus area for this thesis work.

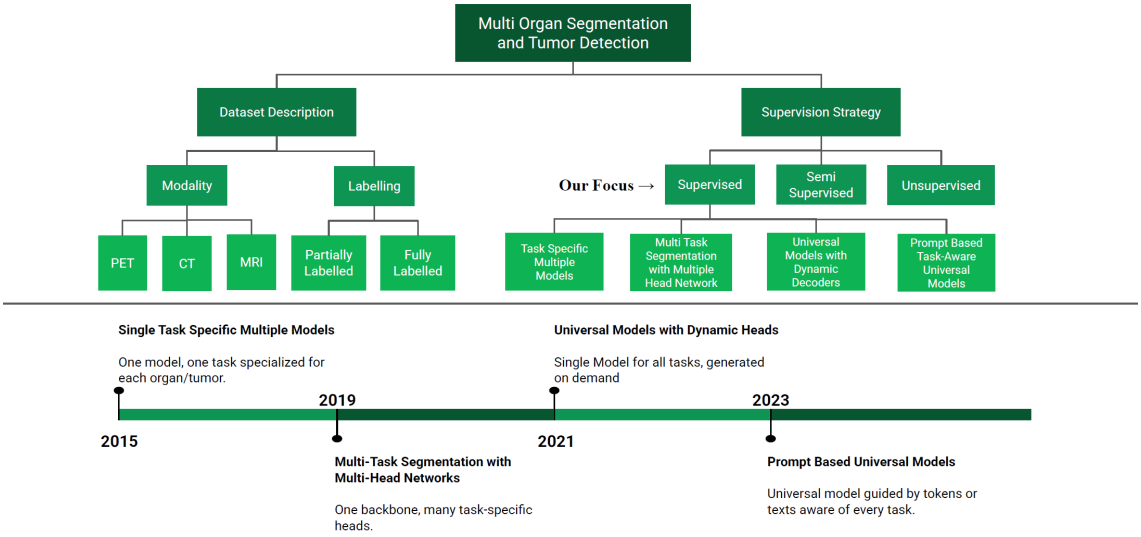

Figure 2.1: Evolution of Multi-Organ Tumor Segmentation and Our Focus Area in this particular domain

2.1 Foundations

The emergence of U-Net [8] revolutionized biomedical image segmentation by introducing a contracting-expanding encoder-decoder architecture with skip-connections. U-Net's capability to produce dense per-pixel predictions even from a few annotated samples laid the groundwork for the first generation of medical segmentation deep learning networks.

Standing on U-Net's shoulders, nnU-Net [5] offered a self-configuring framework that automated hyperparameter tuning, cross-validation, network adaptation, and pre/post-processing, achieving strong performance across diverse biomedical datasets without requiring much manual intervention.

However, these models were inherently designed for single tasks. Each new organ or tumor type would mean having to train a separate model, leading to scalability issues. Furthermore, convolutional architectures, while powerful, struggled to model long-range spatial dependencies critical for complex anatomical relationships. [1]

To address these limitations, TransAttUNet [1] introduced Transformer-based self-attention mechanisms into the U-Net structure, aiming to overcome the local receptive field constraints of standard convolutions. By incorporating both Transformer Self-Attention (TSA) modules and Global Spatial Attention (GSA) modules at the bottleneck of the encoder-decoder framework, TransAttUNet enabled the modeling of non-local dependencies and contextual interactions across the entire image space. Additionally, multi-scale skip connections aggregated semantic features at different resolutions, helping to recover fine anatomical details lost during downsampling. This combination of convolutional localization and Transformer-driven global reasoning significantly improved segmentation performance on complex medical imaging tasks, setting a new direction for hybrid CNN-Transformer architectures in biomedical segmentation.

2.2 Addressing Complexity

To improve scalability, researchers transitioned to multi-task segmentation networks such as Med3D [9] and TAL-Net [10]. Med3D introduced the concept of sharing a common encoder across tasks while utilizing multiple task-specific decoders, enhancing data efficiency. Similarly, TAL-Net adopted adaptive loss functions to manage partially labeled data across multiple decoder heads.

While it did make it easier to train over multiple tasks, these multi-head architectures

incurred high inference costs and lacked flexibility: adding new tasks now requires architectural redesign and retraining. However, managing multiple active decoders simultaneously posed memory and computational challenges.

2.3 Towards Universality

Following the influence of multi-head systems, dynamic segmentation models soon began to come to the limelight. DoDnet [3] first proposed a single encoder-decoder network with a dynamic head that changes on demand, in order to be conditioned on task-specific information. They also managed to encode each task as a one-hot vector and generated dynamic convolution kernels. In addition, Zhang et al. came up with the MOTS dataset by combining various other datasets. DoDNet also enabled training on multiple partially labeled datasets without architectural modifications.

The dynamic approach marked a major shift by addressing the partial labeling issue prevalent in medical datasets. However, the reliance on one-hot encoding for the prompts ignored semantic relationships between tasks, i.e. the natural connection between liver and liver tumor, essentially any organ and its corresponding tumor. Additionally, introducing task-awareness only at the final decoding state limited the model's ability to handle complex anatomical targets effectively.

2.4 Introducing Semantics

To overcome the semantic limitations of one-hot encoding, language-vision models were introduced. The CLIP-Driven Universal Model [4] incorporates text embeddings served from CLIP to represent organs and tumors, capturing anatomical relationships. By using the semantic embeddings instead of rigid task vectors, the model improved generalization and allowed easy extensibility: new tasks could be added simply by providing new textual labels without retraining.

Although performance heavily depended on the alignment between CLIP's pretraining and medical domain knowledge which was a lacking in case of CLIP.

2.5 Early Task Awareness

Building upon the works of Zhang et al. and Liu et al., UniSeg [6] introduced the novel prompt-driven segmentation paradigm. Instead of introducing task-specific conditioning at the decoder output, UniSeg injected learnable universal prompts early into

the model pipeline through a FUSE module, allowing the encoder and decoder to co-train in a task aware manner from the beginning.

The design allowed better modeling of inter-task relationships and improved feature representations across multiple modalities and domains. UniSeg outperformed earlier universal models like DoDNet [3] and CLIP-driven frameworks [4], demonstrating superior Dice scores on various organ and tumor segmentation benchmarks. Nonetheless, UniSeg's prompts were purely learnable without explicit biomedical domain grounding, leaving room for further improvements in semantic fidelity.

2.6 Identified Gaps and Research Opportunities

While significant progress has been made, several gaps persist:

- **Long-Range Spatial Reasoning**: CNN-based backbones like U-Net and nnU-Net have limitations in modeling long-distance anatomical dependencies
- **Semantic Grounding**: Existing prompt-based models lack explicit incorporation of medical knowledge, relying on either learnable or general-purpose language embeddings
- **Cross-Domain Generalization**: Domain shifts across modalities (e.g., CT vs MRI) still challenge universal models, as observed in generalization studies
- **Explainability**: Prompt-based fusion mechanisms remain opaque, hindering clinical trust and interoperability

These gaps motivate our proposed enhancements that aim to replace CNN-based encoders with Transformer-based architectures for better global spatial reasoning, and to enrich task prompts with biomedical semantics via models like BioBERT.

2.7 Summary

In summary, the trajectory of research in multi-organ segmentation and tumor detection has evolved from specialized task-specific models to dynamic universal frameworks, and further towards prompt-driven, semantically informed architectures. Each advancement has addressed critical limitations of its predecessors but introduced new challenges, particularly around semantic understanding and spatial reasoning. Our work builds on these insights to further enhance universality, and domain generalization in medical image segmentation.

Chapter 3

Proposed Methodology

3.1 Introduction

This chapter presents the detailed methodology developed to address the challenges in universal multi-organ and tumor segmentation. Building upon the strengths of convolutional feature extraction, global Transformer-based context modeling, and semantic prompt-driven guidance, the proposed hybrid framework introduces a BioBERT-enhanced FUSE module, integrated into a TransAttUNet-inspired encoder-decoder backbone. By merging fine-grained spatial features with long-range dependencies and biomedical task-specific knowledge, the architecture aims to deliver accurate and generalizable segmentation across diverse modalities.

3.2 Overview of the Methodology

The proposed system follows an encoder-decoder paradigm enriched with global self-attention, prompt-driven task conditioning, and multiscale feature fusion. Specifically, a CNN-Transformer hybrid encoder extracts hierarchical features from the input images. BioBERT embeddings, representing task semantics, are fused with visual features through an enhanced FUSE module to create task-specific prompts. These prompts interact with encoded visual tokens via a transformer decoder module inspired by DETR-like architectures. Finally, a CNN-based decoder reconstructs high-resolution segmentation maps, leveraging skip connections at multiple scales to preserve fine details.

Figure 3.1 illustrates the overall architecture.

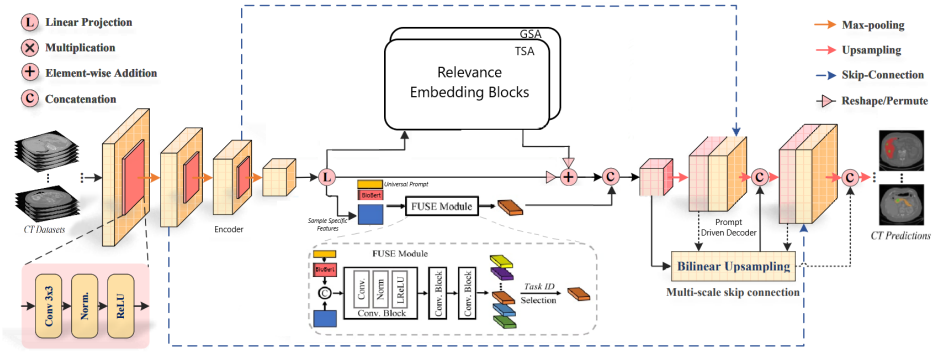

Figure 3.1: Proposed Hybrid Architecture Integrating BioBERT-FUSE Prompting with TransAttUNet-style CNN-Transformer Backbone.

3.3 Chronological Breakdown of Components

3.3.1 Problem Formulation

Given N segmentation tasks $\{D_1, D_2, \ldots, D_N\}$ where $D_i = \{(X_{ij}, Y_{ij})\}_{j=1}^{n_i}$, with X_{ij} representing input images and Y_{ij} their corresponding ground truths, the objective is to learn a unified model $\mathcal{M}$ capable of segmenting all tasks with minimal performance degradation. Straightforward solutions involve training N separate models; however, they ignore task correlations and impose significant redundancy [6].

3.3.2 CNN-Transformer Hybrid Encoder

The encoder integrates convolutional blocks for local feature extraction and Transformer blocks for modeling global context [7], [8]. Each input image $X \in \mathbb{R}^{C\times H\times W}$ passes through:

- **Convolutional stages**: Multiple 3×3 convolutional layers, instance normalization, ReLU activations, and strided down-sampling.
- **Transformer stages**: Vision Transformer layers operating on flattened feature maps, incorporating Multi-Head Self-Attention (MHSA).

MHSA operation at each layer is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.1}$$

where Q, K, V are queries, keys, and values projected from the input features.

Additionally, inspired by TransAttUNet [1], self-attention layers are enhanced with:

Transformer Self-Attention (TSA):

$$TSA(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3.2}$$

Global Spatial Attention (GSA) mechanism aggregates features spatially to complement TSA:

$$GSA(M,N,W_p) = (W_p B)q, \tag{3.3}$$

where M, N are linear projections of features, and W_p learns spatial relationships.

3.3.3 Universal Prompt with BioBERT Embedding

To introduce task-awareness, we design a **universal prompt** embedding module enhanced with BioBERT embeddings, denoted as $\mathrm{F}_{\text{uni}} \in \mathbb{R}^{N\times\frac{C}{8}\times\frac{H}{8}\times\frac{W}{8}}$. The BioBERT vectors capture semantic relations among tasks, enriching the prompts with domain knowledge.

Following UniSeg [6], the prompt fusion is computed as:

$$\{F_{task_1}, \ldots, F_{task_N}\} = \text{Split}\left(f\left(\text{cat}(F_{\text{uni}}, F)\right)\right)^N, \tag{3.4}$$

where F is the extracted feature map from the encoder, $f(\cdot)$ denotes feedforward projection, and $\text{cat}(\cdot)$ represents concatenation.

3.3.4 FUSE Prompt Decoder

The FUSE module concatenates universal and visual features to generate task-specific prompts. Using cross-attention, each prompt interacts with the encoder tokens as:

$$\text{CrossAttention}(P,E) = \text{softmax}\left(\frac{PW_Q(EW_K)^T}{\sqrt{d_k}}\right)EW_V, \tag{3.5}$$

where P denotes prompt queries, E the encoded visual tokens, and W_Q, W_K, W_V are learnable projections.

3.3.5 Multi-Scale Skip Connections

Inspired by TransAttUNet's dense skip connections [1], encoder features at different resolutions are concatenated during decoding:

$$F = f_n(\upsilon_1(F_1) \oplus \upsilon_2(F_2) \oplus ... \oplus \upsilon_n(F_n)), \tag{3.6}$$

where $\upsilon_i(\cdot)$ are upsampling operators and $\oplus$ denotes concatenation.

This dense fusion ensures preservation of fine anatomical details during segmentation.

3.3.6 Segmentation Head

The upsampled feature maps are passed through convolutional refinement blocks, culminating in a 1×1 convolution for final class prediction. For binary segmentation, a sigmoid activation is used; for multi-class segmentation, softmax is applied.

The loss function combines Dice loss and cross-entropy loss as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{Dice}} + \beta\mathcal{L}_{\text{CE}}, \tag{3.7}$$

where α and β balance the contributions (empirically set to 0.5).

3.4 Integration and Data Flow

The data flow is summarized as follows:

1. Image input is processed by the CNN-Transformer encoder.
2. BioBERT-augmented universal prompt is fused with encoder features via FUSE.
3. Task-specific prompts attend to encoded visual tokens.
4. Features are progressively upsampled using skip connections.
5. Final segmentation mask is generated by 1×1 convolution.

Figure 3.1 depicts these interactions visually.

3.5 Summary

The proposed methodology introduces a novel integration of convolutional, Transformer-based, and prompt-driven mechanisms, enhanced with domain-specific semantic embeddings from BioBERT. By combining local and global feature modeling, task-aware prompting, and dense multi-scale feature aggregation, the framework establishes a powerful and extensible approach for universal multi-organ and tumor segmentation.

Chapter 4

Dataset Discussion

We are using three distinct datasets - **MOTS**[3], **BraTS 2021**[11] and **Prostate MRI**[12] to evaluate the performance of multi-organ tumor segmentation models across different modalities and anatomical structures.

4.1 Datasets

- **MOTS:** The MOTS dataset provides a benchmark for multi-organ and tumor segmentation tasks in abdominal CT scans. It contains volumetric CT images annotated with a variety of organ and tumor masks in the abdominal region. It was assembled by combining multiple other datasets(e.g., LiTS, KiTS, MSD) by Zhang et al[3]. However, the dataset shows significant interclass imbalance. Certain organs and tumors are overrepresented while others have relatively few labeled instances. For example, large organs like liver and kidneys are heavily represented but smaller organs such as the adrenal glands have comparatively few labeled examples. Despite having the imbalance, the MOTS dataset captures a wide anatomical range, which introduces both opportunities for robust learning and challenges due to difference in tumor size, location, and morphology.
- **BraTS 2021:** The BraTS 2021 dataset focuses on the segmentation of brain tumors using multimodal MRI scans that include T1, T1Gd, T2, and FLAIR sequences. Each case is annotated with labels corresponding to enhancing tumor, the core of tumor, and whole tumor regions. BraTS segmentation demands the model to learn from multimodal inputs simultaneously and to recognize subtle intensity variations associated with different tumor subregions unlike standard

segmentation tasks. Although the dataset is relatively well-curated and balanced compared to others, challenges such as inter-patient variability, noises, and the fine granularity required for accurate subregion delineation make it a perfect benchmark for evaluating the capability of prompt-guided segmentation models in complex scenarios.

- **Prostate MRI:** The Prostate MRI dataset provides axial T2-weighted MRI scans for prostate segmentation. The dataset is smaller compared to MOTS and BraTS and often demonstrates high inter-institutional variability due to differences in scanner types, imaging protocols, and annotation styles of different institutions. The dataset is made up of healthy and pathological prostate cases, offering a mix of normal and abnormal variations in prostate anatomy. One of the significant challenges with the Prostate MRI dataset is the low contrast between the prostate gland and surrounding tissues, making it a non-trivial segmentation task even for strong models. Class imbalance is not an issue here compared to other datasets, but limited size of the dataset requires careful data augmentation and cross-validation strategies to ensure correct performance evaluation.

4.2 Conclusion of the Chapter

In this section, we introduced and described the datasets we are using in our thesis: MOTS, BraTS2021, and ProstateMRI. Each dataset offers their own set of difficulties and opportunities for driving multi-organ and tumor segmentation tasks across different imaging modalities and clinical setups. A common problem that arises in these datasets is the presence of significant interclass imbalance, where larger anatomical structures, such as the liver and kidneys, are represented more than smaller organs or tumors, such as adrenal glands or small metastases. This imbalance introduces inherent difficulties for model training which often leads to biased predictions favoring the dominant classes as seen in the papers that worked on them.

Understanding the characteristics and limitations of these datasets is needed, as they directly impact model performance and generalizability. The insights we gained here emphasize the need for careful handling of data distribution issues. This sets the stage for evaluating the effectiveness of our proposed pipeline, and the results in subsequent sections will be interpreted in light of these dataset-specific characteristics.

Chapter 5

Conclusion

The primary goal of the research was to understand the existing technology, attempt to identify and address key challenges in the domain of multi-organ and tumor segmentation.

Two core research questions were identified to guide the research:

- **RQ1:** How can we effectively capture long-range dependencies in multi-organ and tumor segmentation tasks?
- **RQ2:** How can task prompts be enriched with domain-specific semantics for improved task generalization?

Answering these questions directed the exploration, analysis, and design of a novel methodology intended to bridge the identified gaps in existing segmentation models.

5.1 Contributions of the Research

To address the Research questions, A hybrid pipeline was proposed that integrate TransAttU-net [1], a Transformer based encoder, with a BioBERT-driven prompt decoder. This architecture was designed to simultaneously address the need for capturing global spatial relationships and incorporating domain-specific semantic information into the segmentation process.

Through a analysis of existing universal segmentation models, such as DoDNet [3], CLIP [4], and UniSeg [6], critical limitations were identified, particularly related to deplayed prompt fusion, insufficient spatial modeling, and domain-specific semantic grounding.

5.2 Future Work

The full implementation and validation of the proposed pipeline, after rigorous testing across diverse datasets, will be necessary to confirm improvements in segmentation and generalization capabilities.

Given the critical nature of medical applications, enhancing model interpretability is an important next step. The integration of explainability modules, such as attention heatmaps and prompt influence visualizations, will make model outputs more transparent, trustworthy, and reliable. This method may be explored in future research.

In conclusion, this research work provides a firm step in the development of a universal model that can seamlessly bridges the semantic and spatial reasoning in medical image analysis. The findings and proposed solutions lay a strong foundation for further advances in the field.

References

[1] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 55–66, 2024. DOI: 10.1109/TETCI.2023.3309626.

[2] X. Liu, L. Qu, Z. Xie, et al., "Towards more precise automatic analysis: A systematic review of deep learning-based multi-organ segmentation," *BioMed Engineering OnLine*, vol. 23, p. 52, 2024. DOI: 10.1186/s12938-024-01238-8. [Online]. Available: https://doi.org/10.1186/s12938-024-01238-8.

[3] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2021, pp. 1195–1204. DOI: 10.1109/CVPR46437.2021.00125.

[4] J. Liu et al., "Clip-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2023, pp. 21 095–21 107. DOI: 10.1109/ICCV51070.2023.01934.

[5] 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. DOI: 10.1038/s41592-020-01008-z.

[6] Y. Ye, Y. Xie, J. Zhang, Z. Chen, and Y. Xia, "Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023*, Springer, 2023, pp. 508–518. DOI: 10.1007/978-3-031-43898-1_49.

[7] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[8] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[9] S. Chen, K. Ma, and Y. Zheng, *Med3d: Transfer learning for 3d medical image analysis*, 2019. arXiv: 1904.00625 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1904.00625.

[10] J. Shang and X. Fang, "Triaxial low-rank transformer for efficient medical image segmentation," in *Pattern Recognition and Computer Vision*, ser. Lecture Notes in Computer Science, vol. 14342, Springer, 2024, pp. 91–102. DOI: 10.1007/978-981-99-8432-9_8.

[11] U. Baid et al., *The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification*, 2021. arXiv: 2107.02314 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2107.02314.

[12] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1013–1023.

[13] P. Liu, C. Gu, B. Wu, X. Liao, Y. Qian, and G. Chen, "3d multi-organ and tumor segmentation based on re-parameterize diverse experts," *Mathematics*, vol. 11, no. 23, p. 4868, 2023. DOI: 10.3390/math11234868.

[14] Y. Xie, J. Zhang, Y. Xia, and Q. Wu, "Unimiss+: Universal medical self-supervised learning from cross-dimensional unpaired data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 021–10 035, 2024. DOI: 10.1109/TPAMI.2024.3436105.

[15] T. Yang and J. Song, "An automatic brain tumor image segmentation method based on the u-net," in *Proceedings of the 4th IEEE International Conference on Computer and Communications (ICCC)*, IEEE, 2018, pp. 1600–1604. DOI: 10.1109/CompComm.2018.8780595.

[16] M. Kang, C.-M. Ting, F. F. Ting, and R. C. Phan, "Rcs-yolo: A fast and high-accuracy object detector for brain tumor detection," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023*, Springer, 2023, pp. 600–610. DOI: 10.1007/978-3-031-43901-8_57.