UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DEPARTMENT OF INFORMATION ENGINEERING

MASTER'S DEGREE IN COMPUTER ENGINEERING

SELF-SUPERVISED VISION TRANSFORMER AND CAUSAL INFERENCE FOR MRI SEMANTIC SEGMENTATION

Supervisor: Prof.ssa Barbara Di Camillo

Candidate: Lorenzo Riccò

Co-supervisor: Prof. Jeremie Sublime

ACADEMIC YEAR 2024 – 2025

Graduation date 13/10/2025

Abstract

Semantic segmentation is a cornerstone of computer vision and plays a critical role in biomedical imaging, where accurate delineation of anatomical structures and tumors supports diagnosis, treatment planning, and disease monitoring. However, the field faces persistent challenges: annotated datasets are scarce, labels are costly and variable, and images often present low contrast and heterogeneous acquisition conditions. These factors make biomedical image segmentation one of the most demanding applications of computer vision.

Over the past decade, the field has undergone remarkable progress, advancing from deep learning–based approaches with convolutional neural networks such as U-Net, to Vision Transformers and more recently, to self-supervised learning frameworks. Each wave of innovation, often appearing in overlapping phases, has addressed part of the problem, yet the fundamental bottleneck of annotation scarcity has remained largely unresolved.

This thesis investigates whether combining the representational richness of self-supervised Vision Transformers with the bias-reduction capabilities of causal inference can improve unsupervised semantic segmentation in biomedical imaging. Specifically, it adapts the CAUSE framework, a causal approach to unsupervised segmentation, to MR images. The thesis contributes both a comprehensive review of the evolution of semantic segmentation methods and an empirical demonstration that integrating self-supervision with causal inference offers a promising direction for annotation-efficient MRI segmentation.

Sommario

La segmentazione semantica è una pietra miliare della computer vision e svolge un ruolo fondamentale nell'imaging biomedico, in quanto la delineazione accurata delle strutture anatomiche e dei tumori supporta la diagnosi, la pianificazione terapeutica e il monitoraggio delle malattie. Tuttavia, il settore deve affrontare sfide persistenti: i dataset annotati sono scarsi, le etichette sono costose e variabili, e le immagini spesso presentano basso contrasto e condizioni di acquisizione eterogenee. Questi fattori rendono la segmentazione di immagini biomediche una delle applicazioni più impegnative della computer vision.

Negli ultimi dieci anni, il campo ha conosciuto progressi straordinari, passando da approcci basati sul deep learning con reti neurali convoluzionali, come U-Net, ai Vision Transformers e, più recentemente, ai framework di apprendimento auto-supervisionato. Ogni ondata di innovazione, spesso comparsa in fasi sovrapposte, ha affrontato parte del problema, ma il collo di bottiglia fondamentale della scarsità di annotazioni è rimasto in gran parte irrisolto.

Questa tesi indaga se la combinazione della ricchezza rappresentazionale dei Vision Transformers auto-supervisionati con le capacità di riduzione del bias dell'inferenza causale possa migliorare la segmentazione semantica non supervisionata nelle immagini biomediche. In particolare, adatta il framework CAUSE, un approccio causale alla segmentazione non supervisionata, alle immagini RM. La tesi offre sia una revisione completa dell'evoluzione dei metodi di segmentazione semantica, sia una dimostrazione empirica che l'integrazione dell'auto-supervisione con l'inferenza causale rappresenta una direzione promettente per una segmentazione della risonanza magnetica più efficiente in termini di annotazioni.

Contents

1. Introduction *9*

1.1 Context **9**

1.2 Motivation and Objective **9**

1.3 Structure of the Thesis **12**

2. Background *15*

2.1 Semantic Segmentation: Application to Biomedical Imaging **15**

2.2 Domain Characteristics and Challenges **18**

2.3.1 CNNs and U-Net 21

2.3.2 CAM-based methods 24

2.3.3 Vision Transformer 30

2.4 Self-Supervised Learning **34**

2.5 Causal Inference **40**

2.6 Foundational and Multimodal Models **46**

3. Method and Contributions *49*

3.1 CAUSE Overview **49**

3.1.1 Step one: Maximizing Modularity 52

3.1.2. Step two: Enhancing Likelihood of Semantic Groups 55

3.1.3 Evidence and Discussion 56

3.2 Domain Adaptation **58**

3.3 Main Contributions **64**

4. Data and Results *67*

4.1 MRI Images and Datasets **67**

4.2 Evaluation **71**

5. Discussion and Conclusion *74*

5.1 Results Discussion **74**

5.2 Conclusion **76**

1. Introduction

1.1 Context

This master's thesis has been developed during my research stay in Paris at ISEP École d'ingénieurs du numérique (https://www.isep.fr/) under the supervision of Professor Jérémie Sublime and Miss Seoyoung Oh, Ph.D. candidate in the DaSSIP research group (Data Science, Signal and Image Processing). Throughout the project, I had the opportunity to work closely with Miss Oh, whose research focuses on deep learning–based semantic segmentation of brainstem structures in MRI with applications to Amyotrophic Lateral Sclerosis (ALS), in a weakly supervised setting. Her work was a central reference point for the development of this thesis in two complementary ways. On the one hand, it provided me with a practical foundation in data-efficient methods for semantic segmentation, exposing both the potential and the applicability of such methods in clinical neuroimaging. On the other hand, the technical challenges she encountered, including the dependence on high-quality annotations and the intrinsic limitations of CAM-based pipelines, served as a catalyst for my own research direction. Motivated by these open questions, I set out to explore alternative paradigms that could push forward the boundaries of semantic segmentation in biomedical imaging, with a particular focus on MRI.

1.2 Motivation and Objective

Semantic segmentation is one of the cornerstones of modern computer vision, enabling dense, pixel-level understanding of complex scenes. In biomedical imaging, it plays a particularly critical role: the ability to accurately delineate anatomical structures, tumors, or pathological regions underpins diagnosis, treatment planning, and monitoring of disease progression. Despite significant advances, the domain continues to face persistent challenges. Annotated medical datasets are scarce, annotations are costly and subject to inter-observer variability, and medical images often exhibit ambiguous boundaries, low contrast, and high variability across institutions and acquisition devices. These factors have made biomedical image segmentation one of the most challenging yet impactful applications of computer vision.

In this thesis, the focus is specifically on Magnetic Resonance Imaging (MRI), one of the most widely used modalities for brain imaging due to its excellent soft tissue contrast and non-invasive

nature. The implementation is centered on MRI brain tumor segmentation, a clinically significant task where precise delineation of tumor subregions can guide treatment planning, monitor disease progression, and provide valuable prognostic information. The complexity of brain tumors, combined with the variability of MRI acquisition protocols across institutions, makes this application both highly relevant and particularly challenging for automated segmentation algorithms.

Over the past decade, the field has experienced remarkable progress, evolving from classical statistical methods to deep learning–based approaches, and from convolutional neural networks (CNNs) such as U-Net, to Vision Transformers (ViT), and more recently, to self-supervised learning frameworks. Each wave of innovation, often appearing in overlapping moments, has addressed part of the problem: CNNs introduced hierarchical feature learning, CAM-based methods opened the door to weak supervision, and Transformers expanded the modeling of global context. Yet, the fundamental bottleneck of annotation scarcity has remained mostly unresolved.

To address this limitation, this thesis began with an extensive review of state-of-the-art strategies for data-efficient segmentation, looking both within biomedical imaging and across broader computer vision. The objective was to identify approaches that could be meaningfully transferred to the biomedical domain, where labeled data is scarce. From this analysis, two directions stood out as particularly promising. The first is the rise of Vision Transformers trained in a self-supervised manner, which are capable of learning rich, semantically meaningful representations directly from unlabeled data, thereby reducing dependence on costly annotations. The second is the adoption of causal inference frameworks, which offer a principled way to mitigate noise and confounding effects, providing answers to the central unsupervised challenge of "what and how to cluster." The combination of these two paradigms, self-supervised representation learning and causality-driven segmentation, establishes the conceptual foundation for the work presented in this thesis. This observation, strongly reinforced during my collaboration at ISEP, is what motivated the central question of this thesis:

How can the semantic richness of self-supervised Vision Transformers be combined with the bias-reduction capabilities of causal inference frameworks to improve unsupervised semantic segmentation in biomedical imaging, particularly MRI?

The objective of this thesis is therefore to investigate a new strategy for data-efficient MRI semantic segmentation by integrating insights from these two powerful research paradigms. On the one hand, self-supervised Vision Transformers can learn rich and transferable representations from large amounts of unlabeled data, reducing reliance on manual annotations. On the other, causal inference frameworks, and in particular the CAUSE method, provide a principled mechanism to mitigate confounding effects in unsupervised or weakly supervised clustering, resulting in more stable and semantically coherent segmentations.

The contribution of this work is twofold. First, it presents a comprehensive literature study that situates these paradigms within the broader evolution of semantic segmentation, highlighting both their promises and their limitations. Second, it develops and evaluates an adaptation of the CAUSE framework to the biomedical domain, focusing on MRI brain tumor segmentation.

To validate this approach, experiments were conducted on two publicly available datasets: the LGG dataset from Kaggle, which provides 2D slices with binary tumor annotations, and the BraTS 2016–2017 datasets from the Medical Segmentation Decathlon, which contain 3D multi-parametric MRI scans with multi-class annotations of tumor subregions. These two datasets were chosen to progressively evaluate the framework: first on a simpler binary segmentation task, and then on a more complex multi-class, volumetric setting.

Results demonstrate that the choice of backbone plays a crucial role. When paired with CAUSE, DINOv2, trained exclusively on large-scale natural image datasets, displayed a notable capacity to generalize to MRI, achieving moderate performance across both tasks. However, visual inspections revealed that its segmentations often lacked precision, suggesting limitations in transferring concepts from natural to medical domains. In contrast, Swin-UNETR, a Transformer architecture pretrained specifically on medical images, consistently outperformed DINOv2 in both quantitative metrics (mIoU, mAP, and Dice) and qualitative segmentation quality. Notably, Swin-UNETR produced more accurate delineations of tumor boundaries and subregions, better capturing the complex structures inherent in brain MRI data.

The combination of CAUSE with medical-pretrained Transformers thus represents a promising direction toward data-efficient, annotation-scarce MRI segmentation, and sets the stage for future extensions of this work to larger-scale clinical applications.

1.3 Structure of the Thesis

To guide the reader through this exploration, the thesis is organized into five main chapters, each of which builds on the previous to develop a coherent narrative. The Background chapter provides the theoretical and methodological foundations of the work. It begins with a detailed overview of semantic segmentation and its critical applications in biomedical imaging, explaining why pixel-level understanding is essential for diagnosis, treatment planning, and monitoring. It then highlights the specific challenges of the domain, such as the scarcity of annotated data, the ambiguous and often fuzzy nature of anatomical boundaries, the multimodal nature of MRI sequences, and the presence of domain shifts caused by heterogeneous acquisition protocols. From there, the chapter traces the architectural evolution of segmentation models, starting with CNNs and the introduction of U-Net as a turning point, before moving to CAM-based weakly supervised methods, which attempted to reduce the dependence on dense labels, and then to Vision Transformers and their hybrid variants, which expanded the modeling capacity to capture both local details and global dependencies. The chapter closes by examining the paradigm shift toward self-supervision, with particular emphasis on the emergent properties observed in methods such as DINO, DINOv2, and DINOv3, and finally introduces causal inference approaches like CAUSE as a promising complement to purely architectural innovations. To ensure a broad perspective, the Background also briefly reviews other emerging research directions, including foundation and multimodal models, prompting-based segmentation, generative diffusion approaches, and causality-aware methods, thereby situating the work within the wider landscape of state-of-the-art developments.

The Method and Contributions chapter represents the methodological core of the thesis. It begins with a comprehensive review of the CAUSE framework, explaining in detail its two-step causal strategy for unsupervised segmentation: the construction of a discretized mediator that serves as a concept clusterbook, and the subsequent use of concept-wise self-supervised learning to refine pixel embeddings and enforce semantic coherence. This section goes beyond a surface-level description by unpacking the intuition, the causal reasoning, and the mathematical formulation behind CAUSE, clarifying why it provides a principled alternative to traditional unsupervised clustering approaches. Building on this foundation, the chapter then turns to the central contribution of this thesis: the adaptation of CAUSE to the biomedical domain, with a specific focus on MRI brain tumor segmentation. Here, particular attention is paid to the distinctive characteristics of MRI data: its volumetric structure, its multimodal acquisition, and the scarcity of reliable annotations,

which make it a particularly suitable but challenging testbed for weakly supervised approaches. The chapter describes in detail the preprocessing pipeline required to handle 3D inputs, including normalization, resampling, and augmentation strategies, as well as the modifications introduced to backbone architectures to make them compatible with volumetric medical data. Special emphasis is placed on the integration of self-supervised Vision Transformer backbones such as Swin-UNETR, which provide semantically rich features well suited to the CAUSE pipeline.

The Dataset and Results chapter presents the experimental setup, including a justification for the use of two complementary public datasets (LGG and BraTS 2016/2017), the preprocessing applied to each, the training configurations, and the evaluation metrics used to assess segmentation quality in both binary and multi-class scenarios experimental outcomes obtained by adapting the CAUSE framework to biomedical imaging, first with DINOv2 backbones and then with Swin-UNETR, evaluated on both the LGG Kaggle dataset (2D, binary segmentation) and the BraTS 2016–2017 datasets (3D, multi-class segmentation). Results are analyzed across multiple parameter settings, including feature dimension, number of clusters in the concept book, and segmentation head variations (CAUSE-MLP vs. CAUSE-TR). Performance is reported using mIoU, mAP, and Dice coefficient, with both quantitative metrics and qualitative visualizations of tumor segmentation quality.

Finally, to contextualize these findings, the Discussion and Conclsuion chapter benchmarks results against the limited body of unsupervised segmentation work available for LGG and BraTS, while also comparing to semi-supervised baselines such as SemiSAM+, which highlight the performance gap between unsupervised and annotation-efficient methods. The discussion emphasizes how general-purpose self-supervised backbones like DINOv2 show encouraging transfer capabilities to medical imaging, but domain-specific pretraining, as in Swin-UNETR, yields more stable and clinically plausible results.

Beyond reporting numerical outcomes, the chapter also reflects on generalization challenges due to domain variability, the role of causal mediation in stabilizing unsupervised clustering, and the limitations of the current implementation, including sensitivity to parameter choices and imperfect tumor boundary delineation. These insights are used to outline future research directions, particularly scaling causal–self-supervised approaches, expanding evaluations to additional

modalities, and exploring large-scale foundational models for robust, data-efficient MRI segmentation.

The source code could be found at: https://github.com/LorenzoRiccoCES/CAUSE-MRI

2. Background

Given the interdisciplinary nature of the topic, spanning different themes in computer vision, it is essential to delineate the relevant theoretical background with clarity and precision. The objective of the chapter is to equip the reader with a coherent and structured understanding of the landscape in which this work is situated, enabling a deeper appreciation of both the challenges addressed and the solution proposed throughout the thesis.

Note that, rather than attempting to exhaustively survey the field, the chapter is deliberately selective, concentrating on the pillars that shape the motivation, formulation, and design of the proposed method. It begins with the fundamentals of semantic segmentation in biomedical imaging and its characteristics, then traces the evolution of architectures from Convolutional Neural Networks to Vision Transformers, before turning to the paradigm of self-supervised representation learning and the role of causal modeling. Together, these elements delineate the landscape in which the thesis is situated and prepare the ground for the methodological choices that follow.

2.1 Semantic Segmentation: Application to Biomedical Imaging

Image segmentation is a classic problem in computer vision research and has become a hotspot in the field [1]. The so-called image segmentation refers to the division of an image into several disjointed areas according to features such as grayscale, color, spatial texture, and geometric shapes so that these features show consistency or similarity in the same area, but between different areas shows a clear difference.

Generally, image segmentation can be described by a set theory model. Given an image *I* and a set of similarity constraints, $C_i(i = 1, 2, \ldots, n)$ the objective is to obtain a division of it, namely:

$$U_{x=1}^{N} R_x = I \; ; \quad R_x \cap R_y = \emptyset; \; x \neq y; \; y \in [1, \; N]$$

where R_x satisfies both sets of all pixels in communication similarity constraint $C_i(i = 1,2, \ldots, n)$ the image areas. The same is true for R_y. . *N* is a positive integer not less than *2*, indicating the number of regions after division.

Among its variants, semantic segmentation is the task of assigning a class label to each pixel in an image, effectively partitioning it into semantically meaningful regions. Unlike classification, which assigns a label to an entire image, semantic segmentation constructs a dense map that distinguishes object boundaries and background areas at pixel-level resolution. Therefore, semantic segmentation can be viewed as dense pixel-level classification, critical for applications necessitating detailed scene comprehension.

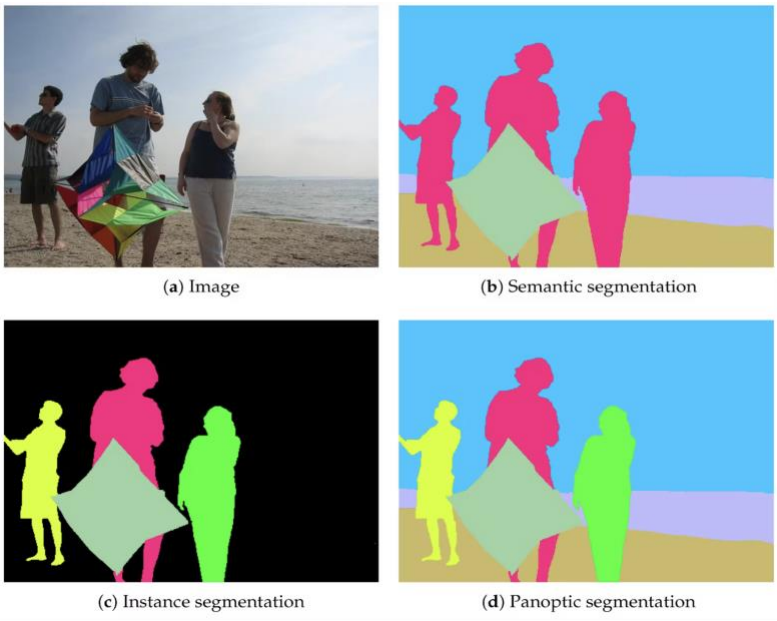


Figure 1: examples of image segmentation techniques. As shown in the figure, semantic segmentation assigns each pixel to a semantically coherent class.

In biomedical imaging, semantic segmentation plays a vital role: imaging data constitute approximately *90%* of healthcare data [2], making segmentation techniques invaluable for medical interventions and diagnostics. Early applications focused on segmenting organs in CT and MRI scans, such as the liver [3], kidneys [4], and brain, facilitating improved diagnosis and treatment planning [5]. In oncology, semantic segmentation has been instrumental in identifying and quantifying tumors, enhancing the accuracy of radiotherapy targeting [6][7]. More recent studies have applied these techniques to segment complex structures like blood vessels and neural pathways, aiding in surgical planning and neurological research [8].

These capabilities are not only theoretical but have been validated in diverse clinical scenarios. For instance, in gastric cancer diagnosis from endoscopic images, segmentation models have achieved a sensitivity of 96% and a Dice agreement of 0.71, substantially improving early detection accuracy

[9]. In histopathology, deep learning–based segmentation has raised average class accuracy for non-melanoma skin cancer from 79% to 83% and overall accuracy from 85% to 94%, enhancing diagnostic reliability [10]. For polyp detection in colonoscopy, modern architectures have achieved Dice scores exceeding 80% on benchmark datasets, enabling earlier and more reliable colorectal cancer screening [11]. These examples illustrate not only the technical progress but also the clinical impact of semantic segmentation.

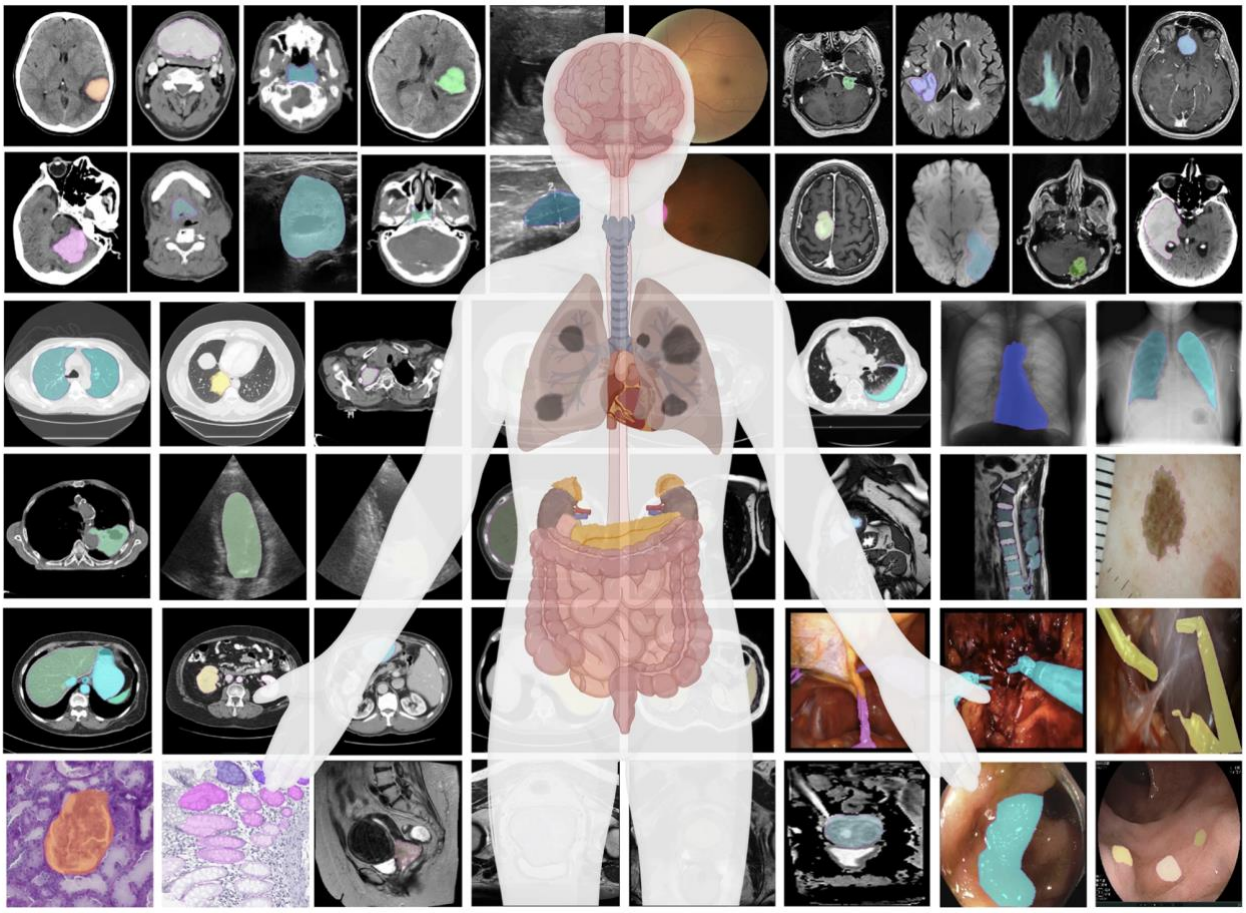


Figure 2: Examples of semantic segmentation in biomedical imaging across diverse modalities and anatomical regions, highlighting its role in delineating organs, tissues, and pathological structures.

The remarkable progress in performance and applicability has been driven by a paradigm shift in methodology over the past decade [12] [13] [14] [15]. Traditional statistical or rule-based approaches, once dominant in the field, have given way to deep learning methods capable of automatically extracting hierarchical features from raw data. The deep learning revolution, fueled by GPU acceleration, introduced Convolutional Neural Networks capable of learning hierarchical features directly from raw image data. Architectures such as U-Net set new benchmarks in biomedical segmentation by combining encoder–decoder pathways with skip connections to

deliver precise predictions even when training data is scarce. More recently, Vision Transformers have expanded the horizon further, modeling images as sequences of patches processed via self-attention to capture global context from the outset, offering advantages in modeling long-range dependencies.

Yet, while these architectural advances have undoubtedly transformed biomedical image segmentation, they have not eliminated the domain's intrinsic difficulties. The scarcity of high-quality annotated datasets, significant variations between imaging modalities, ambiguous or poorly defined anatomical boundaries, and the frequent co-occurrence of multiple organs within the same field of view continue to challenge even the most advanced models. These persistent obstacles remind that progress in architecture must be considered alongside the fundamental characteristics of the biomedical domain, characteristics that continue to shape both current practices and future research directions. It is therefore crucial, before delving into specific architectures, to examine in detail the unique domain characteristics of biomedical imaging, since these define not only the bottlenecks but also the very design choices of segmentation methods.

2.2 Domain Characteristics and Challenges

Peculiarities of the biomedical imaging domain that directly influence how segmentation tasks are approached and solved are a central theme. Biomedical semantic segmentation is not merely a direct application of generic computer vision techniques; the acquisition processes with its clinical requirements and the specific characteristics of the data have driven the development of known architectures. So, by first examining the typical workflow of deep learning–based segmentation models we can better identify where the main bottlenecks arise and how they have shaped the design choices and optimization strategies in the field.

The typical pipeline for deep learning based medical image segmentation begins with data acquisition and dataset preparation. In theory, this stage is straightforward: collect a large dataset and split it into training, validation, and test sets. In practice, it is one of the most significant obstacles in the biomedical domain. Pixel-level annotations require trained specialists (sometimes multiple annotators to ensure reliability) and even then, inter-observer variability can be substantial [16] [17] [18]. Privacy regulations further limit data sharing between institutions, restricting dataset diversity and hindering model generalization. These factors contribute directly to the persistent data labeling scarcity problem in the field.

Once data is available, the next step is preprocessing and augmentation. This involves normalization of intensity values, modality-specific corrections (e.g., bias field correction for MRI), and spatial resampling. Augmentation techniques such as rotations, elastic deformations, and intensity perturbations are crucial to mitigate overfitting and improve robustness. However, in biomedical imaging, augmentations must preserve anatomical plausibility, altering structures too much could make them clinically meaningless. This stage therefore highlights the challenge of domain-specific constraints in data transformation [19].

Furthermore, biomedical images present unique visual challenges. Ambiguous or "soft" boundaries between anatomical structures, low tissue contrast, and the presence of imaging artifacts complicate segmentation techniques in the following steps. Pathological structures may be irregular in shape, vary significantly in size, or blend into surrounding tissues. The co-occurrence of multiple organs or overlapping pathologies within the same field of view adds further complexity, often confusing even expert annotators.

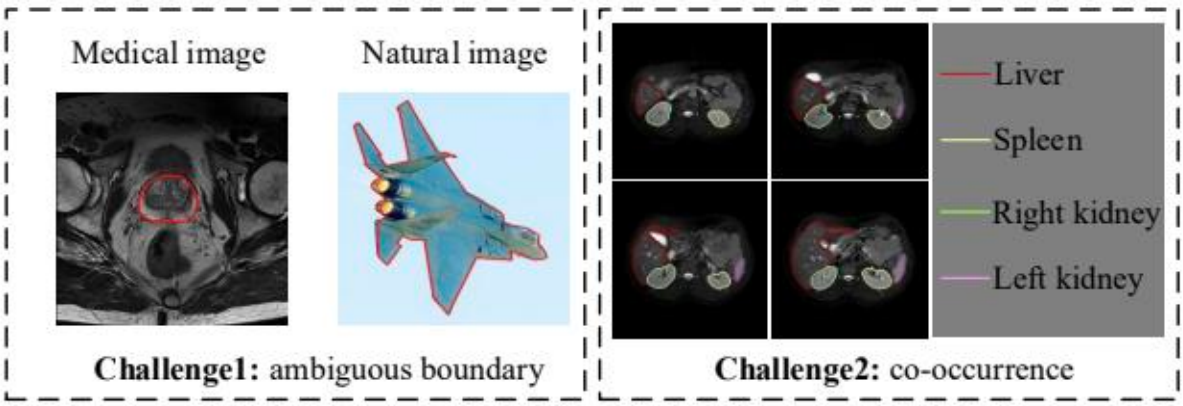


Figure 3: Main challenges of medical images. Challenge1: The object boundary for medical image is more ambiguous than natural image. Challenge2: Different organs often occur in the same medical image in training stage.

The segmentation stage follows, where the choice of architecture, loss function, and optimization strategy is made. Classical CNN-based architectures like U-Net and its variants, as well as more recent transformer-based designs, have all been adapted to biomedical imaging. Yet, these models often contain millions of parameters and require substantial computational resources for training and fine-tuning. This creates an operational and computational barrier for many clinical institutions, especially those lacking advanced hardware or in-house machine learning expertise.

Finally, evaluation and validation of segmentation results is performed using metrics such as the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and Hausdorff Distance (HD). While these metrics capture different aspects of segmentation quality, they are sensitive to the intrinsic characteristics of biomedical images, such as soft tissue boundaries, low contrast, or co-occurring structures, that make accurate segmentation inherently difficult. These visual and structural complexities form the intrinsic image challenges that affect both model training and evaluation.

From this process four persistent and interlinked challenges emerge:

1. **Data-related constraints:** scarcity of large, annotated datasets due to cost, expertise requirements, and privacy concerns.

2. **Generalization and domain shift**: performance drops when applying models to data from different scanners, institutions, or populations.

3. **Computational and operational complexity**: high training and deployment requirements that limit adoption in resource-constrained settings.

4. **Intrinsic image characteristics:** ambiguous boundaries, low contrast, and anatomical variability that complicate accurate segmentation.

In summary, the biomedical imaging domain is characterized by constraints and complexities that influence every stage of the segmentation pipeline. From this process, several interlinked challenges have emerged, each shaping the trajectory of research in biomedical semantic segmentation [16].

2.3 Architectural Evolution

Having established what semantic segmentation is, its relevance in biomedical imaging, and the specific challenges posed by this domain, we now arrive at a key question: *what models have been developed to address these challenges?* The next step is to explore the architectural evolution that has shaped the field of computer vision over the last decade, from the pioneering convolutional

encoder–decoder designs like U-Net to the more recent Vision Transformers. Understanding these architectures and their capabilities will provide the necessary baseline for later introducing the approaches that seek to overcome the domain's persistent bottlenecks.

2.3.1 CNNs and U-Net

Convolutional Neural Networks (CNNs) revolutionized computer vision by introducing architectures capable of automatically learning hierarchical feature representations directly from raw images. Unlike hand-crafted feature engineering, CNNs could detect increasingly abstract visual patterns allowing them to achieve unprecedented performance in image classification tasks. Their success in large-scale benchmarks such as ImageNet [20] established CNNs as the dominant paradigm for visual recognition throughout the 2010s. It was in this context that U-Net emerged.

The U-Net architecture [21] has been a major turning point in biomedical image segmentation, widely regarded as one of the most impactful contributions to the field. Its introduction in 2015 represented a paradigm shift: U-Net enabled accurate, pixel-level segmentation using relatively small, annotated datasets, an essential property in medical imaging.

U-Net follows an encoder–decoder paradigm, where the encoder progressively reduces the spatial resolution while increasing the number of feature channels, and the decoder symmetrically reconstructs the spatial resolution while decreasing the number of channels. Going deeper, the encoder consists of repeated applications of 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling with the feature maps become more abstract as the network goes deeper, capturing high-level semantic information. On the other hand, in the decoder each step in the decoder begins with an upsampling of the feature map, followed by a 2×2 transposed convolution ("up-convolution") that halves the number of feature channels. This is concatenated with the correspondingly cropped feature map from the encoder via skip connections. The concatenated features then pass through 3×3 convolutions and ReLU activations to refine the prediction. The skip connections are a defining feature of U-Net, allowing fine-grained spatial information lost during downsampling to be directly injected into the corresponding upsampling stage improving significantly boundary localization. Finally, a 1×1 convolution maps the multi-channel feature representation at the final decoder layer to the desired number of segmentation classes, producing a dense pixel-level classification map.

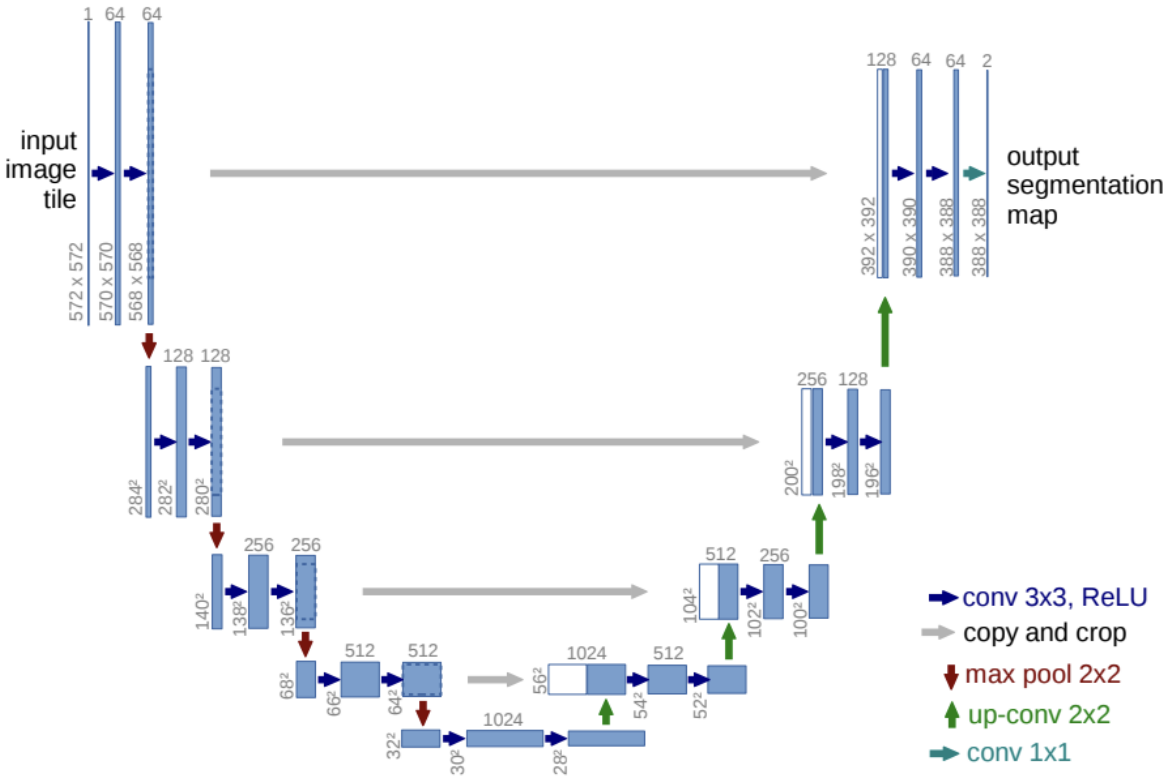


Figure 4: The U-net architecture. Each blue box corresponds to a multi-channel feature map, with the number of channels on top of the box and resolution at the lower left. White boxes represent copied feature maps.

U-Net's architectural simplicity, efficiency, and effectiveness led to the development of numerous variants that further improved segmentation performance. U-Net++ [22] introduced nested and dense skip connections to enhance multi-scale feature fusion, helping to better capture fine-grained structures. Attention U-Net [23] incorporated attention gates that allowed the model to dynamically focus on the most relevant regions of the input. 3D U-Net [24] extended the architecture to handle volumetric data, critical for modalities such as MRI and CT that provide three-dimensional image stacks. nnU-Net [25] further automated the segmentation pipeline, adapting network architecture and preprocessing to specific datasets, and became a strong baseline in medical image segmentation challenges.

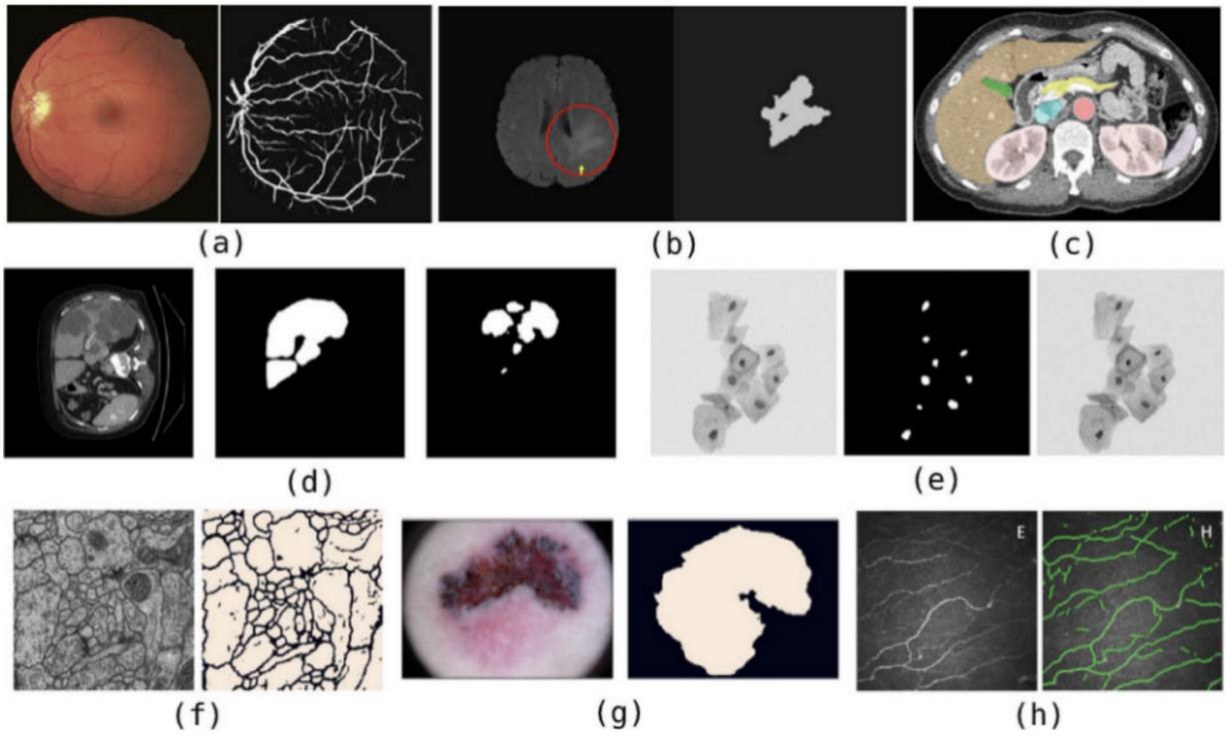


Figure 5: Examples of U-net applications. Images have been collected from (a) retinal vessel segmentation; (b) brain tumor detection and segmentation; (c) multi-organ abdominal segmentation on CT scans; (d) liver tumor segmentation, left to right: original CT image, liver segmentation image, and lesion segmentation image; (e) nuclei prediction, from left to right: original cell images, prediction of nuclei, labeling nuclei in the original images; (f) cell segmentation; (g) skin lesion segmentation; (h) corneal nerve segmentation.

Collectively, these architectures have established U-Net not just as a model, but as a foundational design pattern in biomedical image segmentation. Nevertheless, despite its widespread success, U-Net is not without limitations [13] [26] [27].

One major drawback lies in its limited receptive field: the use of fixed-size convolutional kernels and max-pooling restricts the model's ability to capture long-range dependencies, which are often crucial for recognizing spatially distant but semantically related structures. Although skip connections help restore fine details, the downsampling process inherently loses high-frequency information, and the concatenation strategy cannot always fully recover these fine features, especially for small or thin structures. Additionally, and more importantly, while U-Net performs well with relatively few annotated examples, it still requires pixel-level supervision for training, limiting its practicality in domains like biomedical imaging where annotations are expensive, time-consuming, and often noisy. Although techniques like data augmentation can partially address these issues, they do not eliminate the dependency on expert-annotated labels.

While architectures like U-Net and its numerous variants have set a high standard for biomedical image segmentation, they remain fundamentally dependent on the availability of dense, high-quality annotations. The difficulty and cost of producing such annotations at scale has fueled a paradigm shift toward data-efficient learning strategies, including semi-supervised, few-shot, self-supervised, and weakly supervised semantic segmentation.

2.3.2 CAM-based methods

Data-efficient learning strategies like Weakly Supervised Semantic Segmentation (WSSS) aims to minimize the need for detailed supervision by leveraging coarser forms of annotation, such as image-level labels, bounding boxes, or scribbles. In its most challenging and widely studied form, WSSS with image-level labels provides only a global class annotation for each image, leaving the model to infer object boundaries without explicit pixel-level guidance [28]. Approaches in this domain range from single-stage models that jointly learn classification and segmentation, to two-stage pipelines that first train a classifier to produce coarse localization cues, often through techniques like Class Activation Maps (CAMs), which are then refined into pseudo-masks for training a dedicated segmentation network. This latter family of methods has proven particularly influential, forming a bridge between classification networks and segmentation tasks under weak supervision, and will be the focus of the section.

The original CAM technique [29] was developed as a means of visualizing which regions of an image a classification network focused on when making a decision. This was achieved by replacing the fully connected layers at the end of a CNN with Global Average Pooling (GAP), followed by a linear layer for classification. By backprojecting the weights of the linear layer onto the final convolutional feature maps, CAMs revealed spatial maps indicating the importance of each region for predicting a given class. Overall, CAM provides a way to visualize the areas of an image that are most relevant to a particular class without requiring pixel-level annotations, generating informative localization heatmaps by multiplying the classifier weights with the image features.

Figure 6: Example of CAM, where the heatmap highlights the image regions most relevant to the model's prediction.

In biomedical imaging, CAM-based methods are particularly appealing because they can be trained with image-level diagnostic labels which are far easier and faster to obtain than dense segmentation masks. For example, given an MRI labeled "tumor present," CAM-based techniques can highlight tumor-associated regions, providing a starting point for segmentation without requiring labor-intensive contouring by specialists. Although designed for interpretability, CAMs naturally lent themselves to localization tasks, producing coarse spatial heatmaps that could highlight the position of objects without requiring pixel-level annotations.

So, in the context of WSSS, CAM-based methods adapted classification backbones, typically CNN architectures to produce localization cues from image-level labels. These cues, while often coarse and incomplete, serve as the initial "seeds" for segmentation.

The typical workflow, which is shown in Figure 6, is as follows:

1. Classification and CAM generation
    - Train a multi-label image classifier on the target classes using only image-level labels.
    - Extract CAMs from the trained model for each class present in the image.
    - Threshold the CAMs to produce coarse foreground maps (high-confidence regions) and background maps.
2. Refinement and segmentation:

. Use post-processing techniques, such as Conditional Random Fields (CRFs), random walk, or region growing, to refine the CAM seeds into more complete pseudo-masks.

. Train a supervised segmentation network (often U-Net, DeepLab, or other encoder–decoder architectures) using these pseudo-masks as labels.

This two-stage paradigm allows CAM-based methods to transform weak supervision into a form compatible with standard segmentation pipelines, significantly reducing the need for expensive dense annotations.

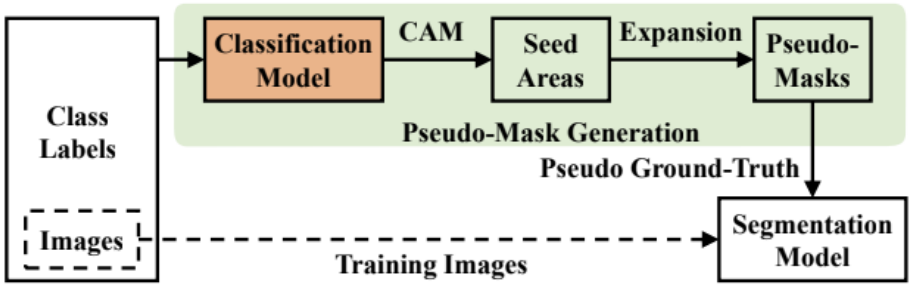


Figure 7: prevailing pipeline for WSSS.

The key technical insight behind CAM lies in the linear combination of feature maps from the final convolutional layer, weighted by the classifier's learned weights for a given class. Formally, for a class *c*, the activation map $M_c(x, y)$ at spatial location *(x,y)* is computed as:

$$M_c(x, y) = \sum w_k^c f_k(x, y)$$

where $f_k(x, y)$ is the activation of the *k-th* feature map at spatial location (x,y) and w_k^c is the weight associated with feature map k for class c in the classifier's final linear layer. After normalization and upsampling to the input resolution, produces a heatmap of class-discriminative regions.

The main drawback of CAM-based methods is that they require architectural modifications, such as the addition of a global average pooling layer. Among these, gradient-base approaches have become the most widely used in medical imaging. Gradient-weighted Class Activation Mapping

(Grad-CAM) [30] was the first to incorporate gradient information into CAM, computing the gradients of the target class with respect to feature map to highlight important regins without necessitating changes to the model architecture Grad-CAM also introduced the guided backpropagation technique, aimed at visualizing which input pixels most strongly activate the network. This advancement spurred multiple extensions, including Grad-CAM++ [31], which improves handling of overlapping objects by weighting gradients according to pixel contributions, and Smooth Grad-CAM++ [32], which enhances robustness and heatmap smoothness by injecting noise into the input data.

In parallel with my research, ongoing work within my group at ISEP has explored CAM-based weakly supervised semantic segmentation for the study of neuroanatomical alterations in Amyotrophic Lateral Sclerosis (ALS), using in vivo MRI. This project addresses one of the central difficulties in biomedical imaging: the scarcity of high-quality voxel-wise annotations. Instead of relying on pixel-level masks, the method uses only image-level diagnostic labels (ALS patient vs. healthy control) to guide segmentation. The approach builds upon a CNN backbone and employs a suite of Class Activation Mapping techniques, including Grad-CAM, ScoreCAM, and LayerCAM, combined with guided backpropagation to enhance fine-grained features. By projecting class-specific activations back to the input space, the method generates heatmaps that highlight brainstem regions most discriminative for the classification task.

The technical pipeline is structured as follows:

1. **Backbone architecture.** A convolutional neural network is trained as a binary classifier (ALS vs. control). The feature maps from the last convolutional layers provide a high-level spatial representation of the input MRI slice.

2. **Class Activation Mapping**. Several CAM techniques are employed to project class-discriminative activations back to the input space. In particular:

    . Grad-CAM uses the gradient of the target class score with respect to feature maps to compute a weighted combination of activations.

- ScoreCAM replaces gradients with class-specific score responses obtained by masking the input with upsampled feature maps, improving robustness to noisy gradients.

- LayerCAM aggregates CAMs computed at multiple intermediate convolutional layers to capture both low- and high-level discriminative cues.

3. **Guided Backpropagation**. To enhance spatial detail, guided backpropagation is used in conjunction with CAMs. This method modifies the backward pass of ReLU units to suppress negative gradients, producing sharper saliency maps.

4. **Heatmap generation and refinement**. The resulting activation maps are thresholded and post-processed to yield binary masks that highlight brainstem subregions most relevant to the ALS classification task. These pseudo-segmentations are then validated against neuroanatomical expectations, providing interpretability and disease-related localization.

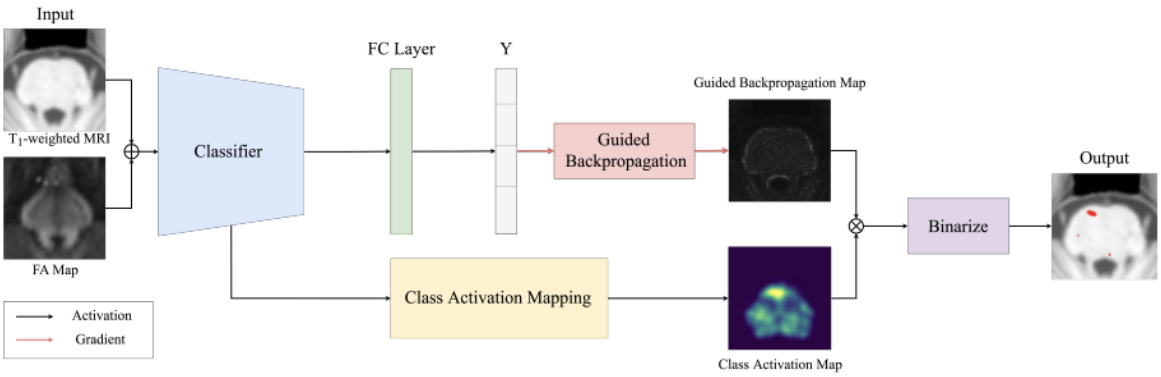
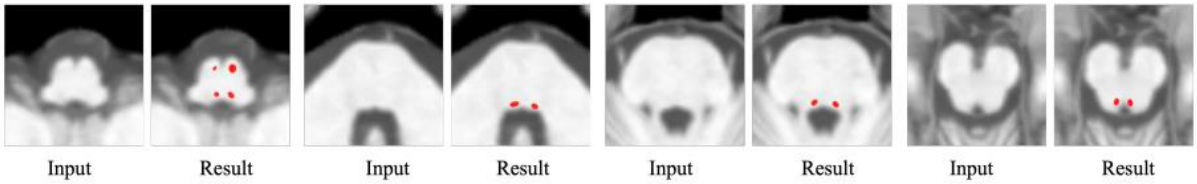


Figure 8: Flow diagram of the WSSS approach. This model takes as input an image composed of a set of slices from T1-weighted MR images and FA maps.

This pipeline demonstrated that even in the absence of pixel-level labels, CAM-based methods can generate meaningful heatmaps that highlight the brainstem regions most associated with ALS pathology. Importantly, the approach was able to capture disease-related localization signals despite the challenges of in vivo MRI, including low resolution, blurred boundaries, and inter-patient variability.



Figutre 9: Examples of segmentations obtained with and LayerCAM with guided backpropagation.

This pipeline demonstrated that even in the absence of pixel-level labels, CAM-based methods can generate meaningful heatmaps that highlight the brainstem regions most associated with ALS pathology. Importantly, the approach was able to capture disease-related localization signals despite the challenges of in vivo MRI, including low resolution, blurred boundaries, and inter-patient variability.

At the same time, the study revealed important limitations. Despite their utility in weakly supervised settings, CAM-based methods have inherent limitations that constrain their effectiveness, particularly in the biomedical imaging domain.

1. **Bias toward discriminative regions**: CAMs tend to highlight only the most class-discriminative areas while ignoring other relevant but less distinctive regions. This results in incomplete object coverage, which can be critical in medical contexts where accurate volumetric or boundary delineation is essential.

2. **Loss of fine-grained detail:** since CAMs are derived from the deepest convolutional layers, their spatial resolution is inherently low. Even after upsampling, the resulting maps often lack precise boundaries, making them unsuitable for direct clinical use without refinement.

3. **Limited global context modeling**: CNN-based CAMs primarily capture local patterns through convolutional filters, which may fail to account for long-range dependencies. In complex biomedical images where structures interact across multiple scales, this limited receptive field can hinder segmentation quality.

4. **Dependence on backbone quality**: the quality of the generated CAMs is directly tied to the classification backbone. If the classification accuracy is suboptimal, the localization maps will also suffer, a frequent challenge when dealing with rare pathologies or highly imbalanced datasets.

From my perspective, this research played a dual role. On the one hand, it provided me with a practical understanding of WSSS techniques and their applicability in clinical neuroimaging. On the other hand, the technical challenges observed in CAM-based segmentation highlighted the need to explore alternative paradigms. If CAM-based approaches demonstrated the promise of weaker supervision, they also made clear that new architectures were needed, ones capable of reasoning globally without sacrificing local detail. It is precisely in this context that Vision Transformers emerged.

2.3.3 Vision Transformer

Vision Transformers (ViT) have rapidly established themselves as one of the most transformative architectures in computer vision, extending the success of Transformer models that had already revolutionized natural language processing. The original Transformer [33], replaced recurrent and convolutional structures with a parallelizable attention-driven mechanism, setting new performance standards in NLP. Inspired by this, Dosovitskiy [34] adapted the architecture to vision tasks, showing that when trained at scale, Transformers could compete with, and eventually surpass, CNNs.

At the heart of the Transformer lies a simple yet powerful encoder block architecture. Each block maintains a constant representation dimension, into which every operation writes its updates. The block is composed of two main stages: multi-head self-attention (MHSA), which allows tokens to interact and exchange information, and a feed-forward network (MLP) that refines each token's representation individually. Both stages are wrapped in residual connections and layer normalization, ensuring stability during deep training. By stacking these blocks, the Transformer builds progressively richer and more contextualized representations, while the residual stream preserves continuity across layers.

The key innovation is the self-attention mechanism. Instead of processing local neighborhoods as CNNs do, self-attention computes pairwise relationships across the entire input sequence. Each input token is linearly projected into three vectors: a query (Q), a key (K), and a value (V). Attention scores are obtained by computing the similarity between queries and keys, scaled by their dimension:

$$Attention(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This mechanism allows every token to "decide" how much to attend to every other token in the sequence, effectively re-weighting information flow based on semantic similarity. In practice, a multi-head attention setup projects tokens into multiple subspaces, enabling the model to capture diverse types of relationships simultaneously, for example, local texture, global shape, or long-range symmetry. Put more intuitively, self-attention can be imagined as a conversation: each token looks around and asks all other tokens *"how relevant are you to me for understanding the bigger picture?"*. Based on the answers, it selectively integrates information, ensuring that global dependencies are considered from the very first layer.

When adapted to vision, Vision Transformers treat an image not as a grid of pixels but as a sequence of patches (e.g., *16×16* pixels). Each patch is flattened, embedded into a high-dimensional vector space, and enriched with positional encodings to preserve spatial relationships. The resulting sequence is then processed through the Transformer encoder, where self-attention captures relationships between any two image regions, regardless of their spatial distance. Unlike CNNs, which rely on local receptive fields and progressively expand their view of the image through stacked layers, ViTs incorporate global context from the very first layer.

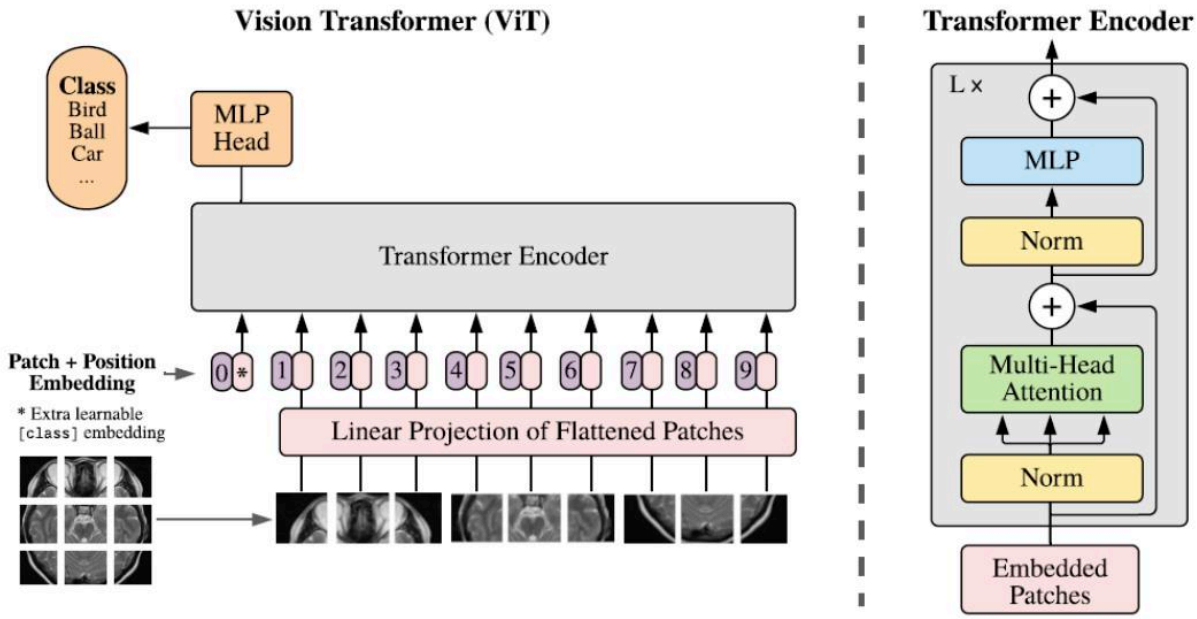


Figure 10: Schematic of the Vision Transformer (ViT). An input image is divided into patches, linearly projected, and combined with positional embeddings (plus a learnable class token). These embeddings are processed by stacked Transformer encoder layers with multi-head self-attention and MLP blocks, enabling global context modeling for classification.

This property has profound implications for biomedical imaging. Anatomical boundaries are often subtle, context-dependent, and span large regions of an image, For example, functionally connected but spatially distant brain areas, or diffuse tumor margins in MRI scans. CNN-based methods, and even CAM-based weakly supervised approaches, often overemphasize the most discriminative local features while neglecting global structure. ViT mitigate this bias: their ability to simultaneously attend to fine-grained details and long-range dependencies makes them especially suited for the complex and multiscale nature of biomedical data. Additionally, because ViT operate on embeddings rather than handcrafted kernels, they can flexibly handle multi-channel or multi-modal inputs (e.g., MRI, CT, histopathology) with minimal architectural modification.

The architectural properties of ViT directly address several weaknesses of CAM-based CNN approaches:

1. **Global context from the start**: Self-attention connects all image patches in the first layer, preventing the overemphasis on only the most discriminative regions. This global context modeling enables more accurate delineation of complex and interconnected anatomical structures compared to CNNs that focus on local features

2. **High-resolution feature preservation**: Patch-based embeddings avoid the repeated pooling and downsampling in CNNs, facilitating finer boundary delineation. For example, techniques like Grouped Multi-Scale Attention (GMSA) apply self-attention at different scales within a single layer and integrate cross-scale feature fusion to improve segmentation precision. These designs help retain high-frequency details required for precise boundary delineation while capturing large-scale context

3. **Flexibility with input modalities**: ViT can ingest multi-channel or multi-modal medical images without major architectural changes, making them adaptable to MRI, CT, or histopathology data.

Building on this, a wave of hybrid architectures has emerged, combining CNNs' strong inductive biases for local feature extraction with Transformers' capacity for global reasoning. Notable examples include TransBTS [35], which fuses 3D CNNs and Transformers for brain tumor segmentation in MRI; CoTr [36], which integrates multi-scale CNN feature maps with deformable Transformers to enhance feature fusion; TransUNet [37], embedding Transformer layers within the U-Net encoder–decoder to outperform state-of-the-art methods in MRI tumor segmentation; and more recent designs like HiFormer [38], H-TUNet [39], and DA-TransUNet [40], which refine cross-scale attention and fusion mechanisms. Across these architectures, the principle is consistent: CNNs capture local structure, while Transformers enrich it with globally coherent representations.

Despite these advances, however, one challenge has persisted across CNNs, CAM-based methods, ViT, and even their hybrids: the dependence on annotated data. Whether supervision is dense, coarse, or weak, segmentation performance remains heavily constrained by the availability and quality of labels. In biomedical imaging, where annotation is expensive, time-consuming, and often subjective, this bottleneck limits the scalability of even the most advanced architectures.

This recognition has shifted the research landscape toward a new paradigm: self-supervised learning. Unlike supervised or weakly supervised methods, which require some form of human-provided labels, self-supervised approaches learn directly from the raw structure of the data itself, uncovering semantic relationships without any manual annotation. When applied to Vision Transformers, self-supervised pretraining has shown exceptional promise, producing general-purpose representations that transfer effectively to downstream segmentation tasks with minimal

annotation. This paradigm and its potential to overcome the annotation bottleneck forms the natural continuation of the architectural journey described so far and stands at the core of the framework explored in this thesis.

In the following section, we explore the principles behind self-supervised learning for Vision Transformers, trace the evolution of these techniques, and highlight the emergent semantic properties that make them uniquely suited to address the challenges at the heart of this thesis.

2.4 Self-Supervised Learning

Self-supervised learning represents a pivotal shift in machine learning, not a mere incremental tweak, but a fundamental rethinking of how models can learn from data [41]. Unlike traditional supervision, which relies on human-annotated labels, self-supervised learning uses the data itself to generate supervisory signals. By designing *pretext tasks*, simple yet meaningful objectives derived from the unlabeled inputs, the model learns to extract rich, generalizable representations without any manual annotation. In other words, pretext tasks act as a mechanism to train models on unlabeled data by defining an artificial objective that encourages the model to learn meaningful representations. For example, a common pretext task involves modifying an image (e.g., rotating it) and training the model to predict the modification (e.g., the rotation angle). By solving such tasks, the model learns features that capture underlying patterns in the data, which can later be fine-tuned for specific downstream tasks like classification or detection. The language domain, by its discrete nature, offers straightforward ways to set up such tasks, which led to many successful unsupervised pre-training approaches for text data with examples including word embeddings, sentence representations, and plain language models. In contrast, computer vision presents greater challenges due to the continuous nature of the signal.

From the early days of self-supervised representation learning, progress has followed a trajectory of increasingly ambitious objectives and architectures, with Vision Transformers becoming a natural match for SSL [42]. The most striking leap came with DINO (Distillation with No Labels) [43], which introduced a self-distillation framework specifically tailored to Vision Transformers. Self-distillation methods rely on a simple mechanism: feeding two different views to two encoders, and mapping one to the other by means of a predictor. To prevent the encoders from collapsing by predicting a constant for any input, various techniques are employed.

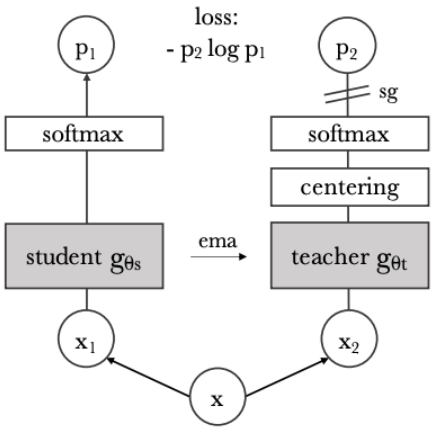


Figure 11: self-distillation with no labels. Two random transformations of the same image are fed to student and teacher networks, which share the same architecture but differ in parameters. The teacher output is batch-centered, and both outputs are normalized with a temperature-scaled softmax. Similarity is enforced via cross-entropy loss, with gradients flowing only through the student (stop-gradient on the teacher). Teacher parameters are updated as an exponential moving average (EMA) of the student parameters.

In DINO, a student network is trained to match the output distribution of a teacher network, updated via momentum, across multiple augmented views of the same image. The student network is trained via standard backpropagation, while the teacher network is updated as an exponential moving average (EMA) of the student's weights, ensuring more stable target representations. The model is trained using multiple augmented views of the same image: the student receives both global and local crops of the input image; the teacher sees only the global views, which helps prevent trivial solutions. The training objective is to align the student's output to the teacher's, across views, using a cross-entropy loss. This setup encourages the student to produce consistent and invariant representations regardless of the image crop, scale, or other augmentations. Overall, DINO trains the student to predict the teacher's distribution over features across views.

The true revelation came in the properties observed in DINO-trained ViT: without ever being told what an object is or where it lies, the self-attention heads in deeper layers spontaneously concentrated on semantically coherent regions, often tracing object boundaries with surprising precision. These attention maps, discovered in the original DINO paper, were not engineered outputs but natural by-products of the SSL process, suggesting that the architecture was internalizing a form of spatial and semantic grouping as a structural prior. For segmentation, this is transformative: it means that the model itself can provide a strong starting point for object

localization and segmentation, effectively generating *free supervision* from its own representation space.

Figure 12: Image taken from the original paper. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

The unexpected discovery that a ViT trained solely in a self-supervised manner can automatically isolate and delineate semantically coherent visual concepts revolutionized the field by showing how powerful object-centric semantic understanding could arise naturally from raw visual data.

Building on these insights, DINOv2 [44] took the paradigm further. DINOv2, released in April 2023, builds upon the principles of DINO but introduces several key innovations that enhance its performance, scalability, and applicability. One of the major advancements in DINOv2 is the incorporation of techniques introduced in the iBOT paper. iBOT (Image BERT Pre-Training with Online Tokenizer) [45] is a self-supervised learning method that combines masked image modeling with the distillation approach used in DINO. Inspired by the masked language modeling objective in NLP (e.g., BERT), Masked Image Modeling randomly masks a subset of image patches and tasks the model with reconstructing the missing visual content. For Vision Transformers, this typically involves tokenizing the image into patches, masking a large proportion (often 40–75%), and using the remaining visible patches to predict the missing ones.

This hybrid approach allows the model to learn more robust and generalizable features by reconstructing masked portions of images while maintaining the benefits of knowledge distillation.

The ablation study in the DINOv2 paper highlights several critical improvements over DINO:

- Improved Feature Learning: DINOv2 introduces a more efficient feature learning process that reduces the need for task-specific fine-tuning. This is achieved through a combination of self-supervised learning and lightweight task-specific modules, such as linear classifiers.

- Out-of-Domain Generalization: DINOv2 demonstrates remarkable generalization capabilities, performing well on tasks even when evaluated on datasets outside its training domain. For example, it outperforms specialized state-of-the-art models on depth estimation tasks, both in-domain and out-of-domain.

- Scalability: DINOv2 is trained on a significantly larger scale, both in terms of model size and dataset size. This scalability allows it to learn more comprehensive visual representations, leading to better performance across a wide range of tasks

The result: robust, universal visual features that deliver state-of-the-art performance on segmentation, depth estimation, classification, and retrieval—even with frozen backbones and no fine-tuning.

Following DINO, other methods built on different pretext tasks to refine and diversify the properties of SSL-ViTs. BEiT [46] applied masked image modeling inspired by BERT, predicting discrete visual tokens to encourage context-aware semantics; MAE [47] simplified reconstruction-based learning by masking up to 75% of image patches and reconstructing the missing pixels with a lightweight decoder, producing highly efficient pretraining pipelines; and I-JEPA [48] shifted the objective away from raw pixel reconstruction toward predicting high-level feature embeddings for missing parts of the image, thereby reducing sensitivity to low-level noise and focusing on semantic content. Each approach contributed pieces to the SSL puzzle: richer global context, better local detail preservation, and stronger robustness to domain shifts.

This evolution culminates in DINOv3 [49], Meta AI's most advanced iteration of the DINO family, engineered to push the boundaries of self-supervised dense representation learning at unprecedented scale. While retaining the self-distillation core - a student network learning to match a momentum-updated teacher across multiple augmented views - DINOv3 introduces critical innovations aimed at stability, scalability, and dense prediction quality. A central contribution is Gram matrix anchoring, a regularization technique that aligns the second-order statistics (Gram matrices) of features between teacher and student. This prevents the drift and degradation of dense features that often occur during prolonged training on large, high-capacity Vision Transformers, ensuring that semantic and spatial coherence is preserved across the network's depth. In addition to Gram anchoring, DINOv3 revises the augmentation pipeline and loss balancing to better serve

dense prediction tasks, moving beyond purely classification-oriented pretraining. The architecture is tuned for high-resolution inputs, allowing the model to capture fine-grained spatial details while maintaining global context, an essential property for segmentation and other localization-heavy applications. Crucially, DINOv3 scales up every dimension of the training regime: from 1.1B parameters in DINOv2 to 7B parameters in DINOv3, and from 142 million training images in v2 to an astounding 1.7 billion images in v3. These scale increases, combined with optimized distributed training and memory-efficient batching strategies, enable learning from massive datasets without instability.

Figure 13: High-resolution dense features learned by DINOv3.

The results are striking. DINOv3 consistently produces high-fidelity, semantically structured dense maps that not only preserve the emergent object-awareness first seen in DINO but also exhibit greater consistency and robustness across diverse domains, including satellite images. Evaluations show that DINOv3 outperforms both its predecessors and other state-of-the-art foundation models in a variety of downstream tasks, including semantic segmentation, object detection, and depth

estimation, often without any fine-tuning. In the context of this thesis, DINOv3 represents the culmination of a decade-long trajectory toward self-supervised, semantically rich, and densely transferable visual representations.

While large-scale self-supervised pretraining has demonstrated remarkable success in learning strong, universal visual representations, these features are not always directly optimal when transferred to biomedical imaging. The domain is uniquely demanding as we have already pointed out in the previous chapter: anatomical boundaries are often weak or ambiguous, contrast between tissues is low, and multiple structures frequently co-occur within the same field of view. For downstream tasks like tumor delineation or organ segmentation, these characteristics make fine-grained local detail crucial. And this is true also in other challenging fields. This is the reason why this gap has motivated a parallel line of research aimed at post-hoc refinement of SSL features to improve their utility in dense prediction tasks under such challenging conditions. Approaches include clustering-based fine-tuning [50] [51], patch-sorting objectives [52], and feature alignment strategies that sharpen local coherence [53]. Within this line of work, a particularly impactful framework is STEGO [54], which builds directly on Vision Transformer embeddings from self-supervised models like DINO and adapts them to segmentation through contrastive learning over pixel-level representations.

The core of STEGO's training involves a contrastive loss function that operates on pixel-level features. This loss encourages the model to pull together features from semantically similar regions, both within the same image and across different images and push apart features from semantically dissimilar regions, ensuring clear boundaries between different classes.

Remarkably, STEGO significantly outperformed previous segmentation baselines, achieving state-of-the-art results and demonstrating notably accurate delineation of object boundaries, internal object structure, and semantic coherence. Its effectiveness comes from a combination of elegance and generality; by simply refining representations already learned by self-supervised Vision Transformers, it delivers high-quality segmentations in domains where annotated data is scarce or prohibitively expensive to obtain.

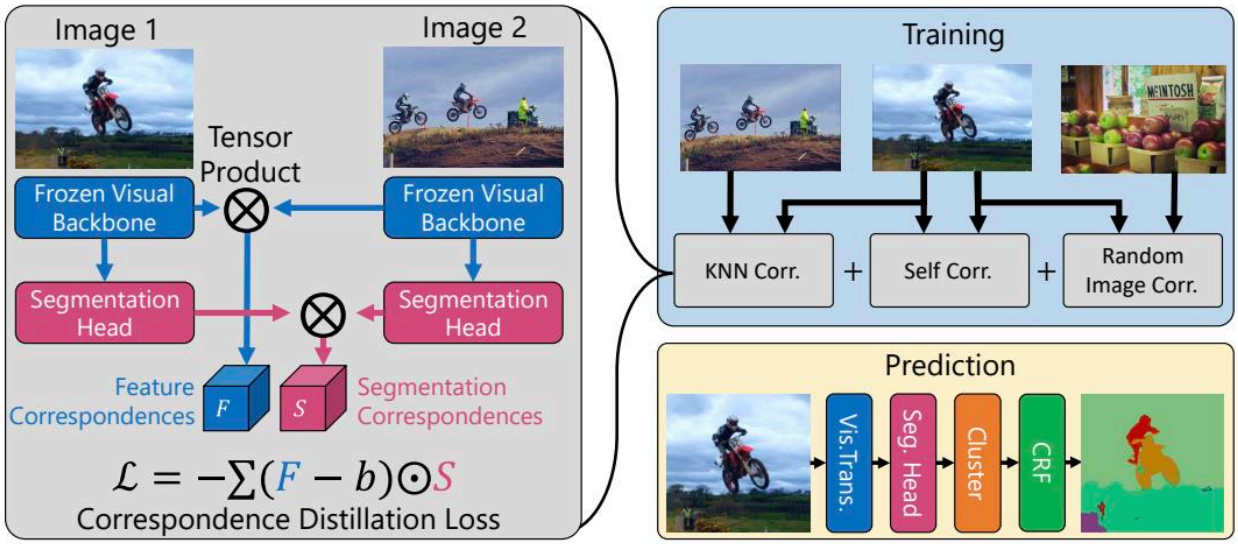


Figure 14: High-level overview of the STEGO architecture during training and prediction. The correspondence distillation loss aligns feature correspondences from a frozen visual backbone with segmentation correspondences, instantiated through KNN, self, and random image correspondences (grey boxes).

This transformative result clearly illustrates how the integration of clustering methodologies and self-supervised ViT embeddings can bridge the gap between implicit representation learning and explicit semantic segmentation. While it does not fully resolve all challenges, STEGO illustrates how self-supervised features can be adapted post hoc into strong domain-specific representations, offering a compelling blueprint for further innovations, among which, causal inference–based frameworks like CAUSE represent a natural next step.

2.5 Causal Inference

Up to this point, we have seen how architectural innovations from CNNs and U-Net, to CAM-based weakly supervised methods, to Vision Transformers, self-supervised learning, and clustering refinements like STEGO have steadily pushed the boundaries of what is possible in semantic segmentation.

Yet despite these advances, another central obstacle remains. Models are often excellent at exploiting correlations in the training data, but far less capable of reasoning about the underlying causal structures that shape the images they analyze. Actually, at the core of science lies the distinction between correlation and causation. The central point is the following: observing that two variables move together does not imply that one causes the other. Classical deep learning

models have been largely correlation-seeking: they excel at exploiting statistical associations in the data, but often without regard for the underlying causal mechanisms.

Typical deep learning-based computer vision tasks can be summarized as follows: Given an image X, the goal is to build a network to predict its label Y correctly [16]. A statistical model fitted with a suitable objective function is often used to estimate the conditional probability distribution $P(Y|X)$. However, only under the independent identical distribution (*I.I.D.*) hypothesis can the learned conditional probability distribution $P(Y|X)$ be applied appropriately from the training set to the testing set. It requires that new prediction samples stay consistent with the distribution of the training set. To minimize the impact of domain differences between two sets, deep learning-based methods such as domain adaptation [17] and generalization [18] have been proposed. However, the problem of generalizability cannot be fundamentally solved through these methods, as the bias caused by domain differences is not fundamentally eliminated. Additionally, since deep learning-based methods benefit from many stacked layers and parameters for approximating high dimensional complex functions, it is hard to interpret the learned models and parameters [16]. The limitation of classical deep learning methods is particularly pronounced in biomedical imaging. The relationships between tissues, organs, or pathological structures are not arbitrary correlations but reflect biological processes. For example, the appearance of a tumor is not just a co-occurrence of intensity patterns, but the manifestation of a pathological process that alters surrounding anatomy. Segmentation models that rely purely on correlations may fail when these superficial cues change, such as with a new scanner, a different patient population, or subtle clinical variations because they never capture the causal factors that truly drive the visual patterns.

In conclusion, existing statistical correlation-based methods overly rely on the given data and analyze the problems through approximating high dimensional functions rather than mechanical mining. There are areas for improvement in both generalizability and interpretability. Due to its mechanical strengths, causality has gained significant attention recently and has rooted its developments across several fields.

Here lies the promise of causal inference: a paradigm that explicitly models cause–effect relationships rather than surface-level correlations. By integrating causal reasoning into vision models, the long-term goal is to develop systems that generalize more robustly, adapt to new environments, and remain reliable even under distribution shifts. For the biomedical domain, where

data scarcity and variability are constants, causal approaches provide a principled way to mitigate biases, disentangle spurious associations, and extract representations that align more closely with the reality of human anatomy and pathology.

Basic causal methods can be divided into two main aspects, causal discovery and causal inference [27]. Based on the causal structures learned by causal discovery, causal inference leverages those relationships for further analysis.

Causal discovery aims to build causal relationships with the structural causal models (SCMs), while causal inference is used to estimate the causal effect. At the center of this framework are Directed Acyclic Graphs (DAGs), which represent variables as nodes and causal influences as directed edges. DAGs are causal structure that allow us to encode hypotheses about how variables interact and to reason systematically about confounding, mediation, and bias.

Based on the causal structure, a structural causal model can specify how each variable is influenced by its parent variables by effect functions. Given a set of variables *X1, X2, ..., Xn* indicating the nodes in the causal structure, they can be written as the outcome of their parent variables and the effect function

$$X_i = f_i\left(PA_i,\ U_i\right),\ i = 1, \ldots\ , n$$

where PA_i indicates the parent variables of X_i , U_i are the unobserved background variables including the noise, and f_i indicates the functional relationship. The purpose of the causal inference to be presented later is to characterize f and estimate the effects of interventions or counterfactuals.

Given a causal structure, causal intervention and counterfactuals can be conducted. The former aims to change the structure by specifying variable data to artificially alter the relationship between variables. In contrast, the latter aims to estimate the variables in a situation opposite to reality. Two typical intervention methods: back-door adjustment and frontdoor adjustment will also be discussed.

Causal intervention refers to the deliberate manipulation of one or more variables in a system to observe the effect on another variable. The foundation of causal inference lies in intervention, actively manipulating one variable to observe its effect on another. Formally, this is described using Pearl's do-calculus: *do(X=x)*. Unlike conditional probability, which only observes correlations, an intervention severs the influence of parent variables on *X* and assigns it directly. In vision tasks, this can be imagined as artificially modifying an input image (e.g., masking a tumor in an MRI scan) and observing whether the model prediction changes. If it does, the region is causally relevant. Interventions therefore provide a way to uncover the true drivers of model predictions.

In many cases, direct intervention is not feasible. Instead, causal inference offers mathematical tools to estimate causal effects from observational data. The backdoor adjustment is one such tool. A backdoor path occurs when a confounder *C* influences both the treatment *X* and the outcome *Y*. In medical imaging, for example, scanner type (*C*) may affect both the input MRI image (*X*) and the annotated diagnosis (*Y*). If left unaccounted for, the model may simply learn scanner-specific patterns.

The backdoor criterion states that by conditioning on *C*, one can block this spurious path and recover the true effect of *X* on *Y*. Mathematically:

$$P\left(Y \mid do(X)\right) = \sum P\left(Y \mid X, c\right)P(c)$$

Where $P(Y \mid do(X))$ represent the causal effect of setting X to some value on outcome Y. The 'do' operator means we intervene on X, not just observe it. This adjustment allows models to disentangle meaningful signal from confounding noise.

Sometimes, confounders are unobservable, making backdoor adjustment impossible. In these cases, frontdoor adjustment comes into play. This method relies on a mediator *M* that lies between *X* and *Y*. By modeling the effect of $X \rightarrow M \rightarrow Y$, it becomes possible to estimate causal effects even without directly observing the confounder.

For example, in vision, a mediator might correspond to intermediate representations extracted by a neural network. These representations can carry the effect of the input while filtering out hidden biases, allowing causal estimation through the $X \rightarrow M \rightarrow Y$.

These tools allow us to move beyond correlation and begin isolating robust cause–effect relationships, even in observational datasets where randomized trials are not possible.

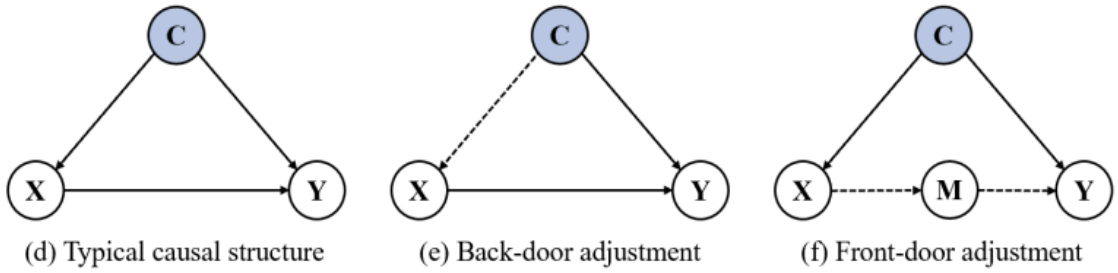


Figure 15: In each graph, X represents a set of input variables, Y represents a set of output variables, and C represents the confounder that causes the spurious correlation. The dashed line represents the block-out path and the blue nodes represent the confounder.

Causal inference has recently entered the machine learning landscape as a powerful complement to correlation-driven methods. Its integration brings three main advantages:

- **De-biasing representations**: By modeling confounding factors explicitly, causal frameworks can enforce representations that are invariant to spurious cues (e.g., scanner-specific noise, demographic bias).
- **Generalization under domain shift**: Causal reasoning helps identify which relationships will remain valid across different datasets and populations, mitigating the "dataset shift" problem that plagues current ML systems.
- **Disentanglement and interpretability**: Causal models naturally aim to separate underlying generative factors, yielding more interpretable representations that clinicians can trust.

In segmentation, these ideas translate into models that not only achieve high accuracy on a benchmark dataset but also remain reliable when deployed across hospitals, scanners, or even pathologies unseen during training.

In recent years, numerous studies have applied causal inference techniques to estimate the true causal effects between treatments and outcomes of interest. The fundamental approach to achieve causal identification involves blocking backdoor paths induced from confounders. In several computer vision methods have employed various causal approaches such as backdoor adjustment establishing explicit confounders [56] [57], mediation analysis [58], and generating counterfactual augmentations for randomized treatment assignments [59] and have been successfully applied at task-specific levels. More recently, various works have demonstrated that the causal approach.

The CAUSE framework (Causal Analysis for Unsupervised Semantic Segmentation) [60] represents one of the first attempts to bring this paradigm into the realm of segmentation. Building on the foundations of self-supervised Vision Transformers, CAUSE integrates causal inference principles into the training pipeline, encouraging the model to identify and exploit causal structures within images. In doing so, it moves beyond correlation-driven clustering and takes a step toward causally grounded representation learning. These principles set the stage for the CAUSE framework (Causal Attention for Unsupervised Semantic Segmentation), which operationalizes causal inference within the domain of semantic segmentation. CAUSE integrates causal reasoning with modern neural architectures, demonstrating how causality can help disentangle true object-level features from spurious correlations in unlabeled or weakly labeled settings.

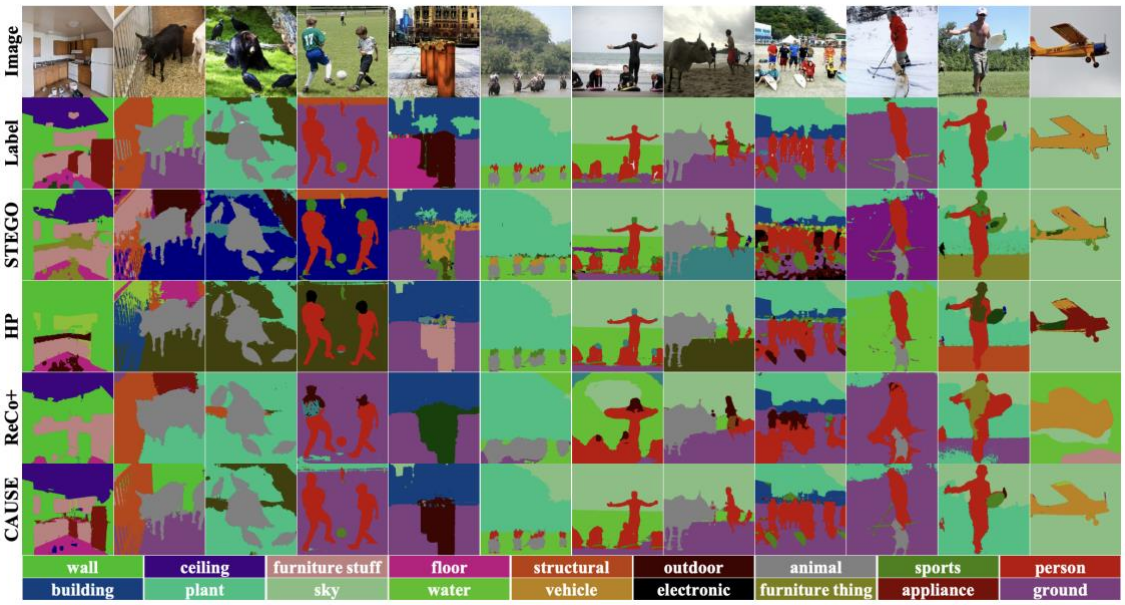


Figure 16: Qualitative comparison of unsupervised semantic segmentation methods. While STEGO, HP, and ReCo+ often produce noisy or incomplete masks, CAUSE achieves more coherent and semantically consistent segmentations.

For this thesis, CAUSE is the final conceptual step before moving into the methodological core. Understanding causal inference allows us to appreciate why CAUSE represents more than just "another architecture": it is an attempt to incorporate cause–effect reasoning into feature learning, with the long-term goal of achieving segmentation models that are not only accurate but also generalizable and trustworthy.

With this, the reader has all the necessary foundations to approach the central research question of this thesis: how can the semantic richness of self-supervised Vision Transformers and the bias-reduction capabilities of causal inference frameworks be jointly leveraged to advance semantic segmentation in biomedical imaging? The next chapter introduces the methodology and architectural design through which this thesis seeks to address this question.

2.6 Foundational and Multimodal Models

This final subsection, before introducing the method pursued, is intended to provide the reader with a broader perspective on ongoing research directions in semantic segmentation, beyond the specific focus of this thesis. While not examined in detail in the following lines, these approaches illustrate how the field is rapidly evolving and highlight possible trajectories for future work. The aim is to

situate the thesis within the wider landscape and to show awareness of complementary paradigms being actively explored.

Before discussing these directions, it is useful to briefly clarify two key terms that frequently appear in recent literature. Foundation models **r**efer to large, pretrained networks trained on massive datasets (often spanning billions of samples), which can then be adapted to a wide range of downstream tasks with minimal task-specific supervision. DINOv2 and DINOv3, cited in the previous chapters, are considered foundation models due to their capacity to perform incredible results in different tasks such as semantic segmentation, object detection and depth estimation among others. Their strength lies in capturing broad, transferable representations rather than being restricted to a single domain. Multimodal models, in turn, are architectures designed to jointly process and integrate data from multiple modalities (most common text and images) enabling tasks such as text-guided segmentation or cross-modal retrieval. In the biomedical domain, the integration of clinical images with textual annotations or reports is an area of growing interest, although challenges such as domain mismatch and limited paired data remain significant obstacles.

Large-scale foundation models such as CLIP [61] and the Segment Anything Model (SAM) [62] have demonstrated unprecedented generalization capabilities in natural images, with performance similar to the latest DINOv3. CLIP couples vision and language representations, enabling zero-shot recognition and segmentation guided by textual prompts, while SAM achieves interactive segmentation through a universal "promptable" architecture. SAM follows the same paradigm of large language models: train once on an enormous dataset, and segment any object in any image with no further training. All it needs is to be prompted by the user via point clicks, bounding boxes, or potentially even text (not released). In practice, SAM is a promptable segmentation machine, giving full control to the user. The scale of the data is unprecedented: SAM was trained on a dataset containing over 1.1 billion masks across 11 million images, the largest segmentation dataset ever built.

Early explorations in biomedical imaging have attempted to fine-tune or adapt these models to radiology, histopathology, and MRI. However, their performance is constrained by domain shift, medical images differ drastically from the natural images on which these models were trained. Moreover, clinical datasets rarely provide the large-scale paired image–text corpora required to fully exploit multimodal pretraining.

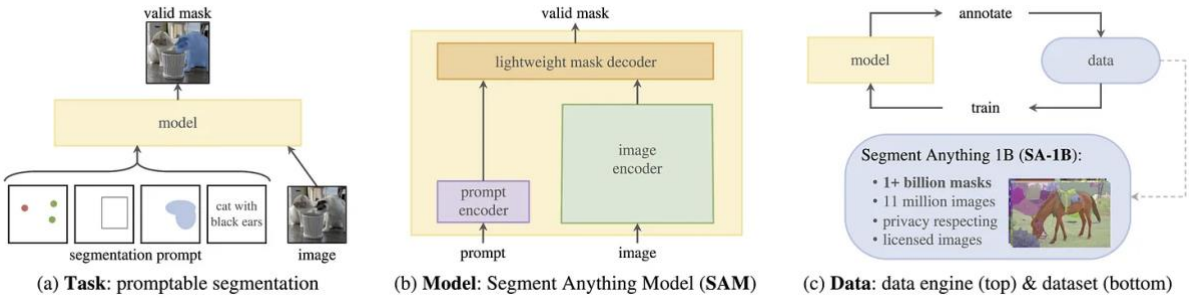


Figure 17: as a foundation model for segmentation, SAM introduces a new promptable segmentation task, a model, and the data engine.

Prompt-driven methods, whether textual, visual, or spatial, are emerging as a flexible alternative to dense supervision. In medical imaging, scribble-based or point-based prompting has been investigated as a practical way to guide segmentation while minimizing annotation cost. For instance, scribble-supervised models have achieved competitive results on cardiac MRI and abdominal CT by using sparse expert input as weak guidance [63]. Similarly, CLIP-inspired models are being adapted to medical imaging to leverage clinical language as prompts. These approaches offer a pragmatic trade-off between human input and automation, although their robustness across different modalities and tasks remains an open challenge.

Generative modeling, particularly diffusion-based frameworks, is beginning to influence segmentation research. Diffusion models can generate detailed pixel-level predictions conditioned on input images, while also capturing uncertainty, an especially relevant property for clinical decision-making. In medical imaging, diffusion has been applied to organ and tumor segmentation in CT and MRI, showing improvements in capturing fine structural details and handling noisy data [64]. Nonetheless, these methods are computationally demanding, require extensive pretraining, and raise questions about interpretability, which currently limit their adoption in clinical workflows.

3. Method and Contributions

The previous chapter traced the evolution of semantic segmentation in biomedical imaging, beginning with its fundamental applications, highlighting the domain-specific challenges, and then following the trajectory of architectural innovations. Along this path, a consistent limitation emerged: the heavy reliance on annotated datasets, which remain scarce, costly, and difficult to obtain in biomedical contexts. This challenge motivated the paradigm shift toward self-supervised learning, where methods like DINO revealed emergent semantic properties that could be leveraged without manual labels. Building on this momentum, we also reviewed new directions, including STEGO for unsupervised segmentation and CAUSE, which introduces causal reasoning as a means of disentangling spurious correlations and improving segmentation quality.

This chapter builds directly on that foundation. Its aim is to detail the methodological framework explored in this thesis: adapting the CAUSE framework to the biomedical domain, with a particular focus on MRI data. While CAUSE was originally developed for natural images, its underlying principles, using causal inference to separate meaningful object–context relationships from confounders, hold promise for addressing the unique challenges of medical imaging, such as subtle anatomical boundaries, high variability across patients, and the limited availability of dense annotations.

Accordingly, this chapter will first outline the CAUSE framework in its original formulation, then describe the modifications and design choices necessary to adapt it to 2D and 3D biomedical imaging. This includes adjustments to data preprocessing, architectural backbones, and training strategies. The objective is to provide a clear and transparent account of how CAUSE was implemented in this domain, setting the stage for the experimental evaluation in the following chapter.

3.1 CAUSE Overview

CAUSE tackles a core pain point in unsupervised semantic segmentation: the clustering target is indeterminate. Without labels, it’s unclear what to cluster (e.g., parts vs. whole objects) and how to group pixels consistently: CAUSE reframes the challenge with a causal diagram and treats

"what/how to cluster" as an unobserved confounder U that opens a backdoor path between pre-trained features T and semantic groups Y.

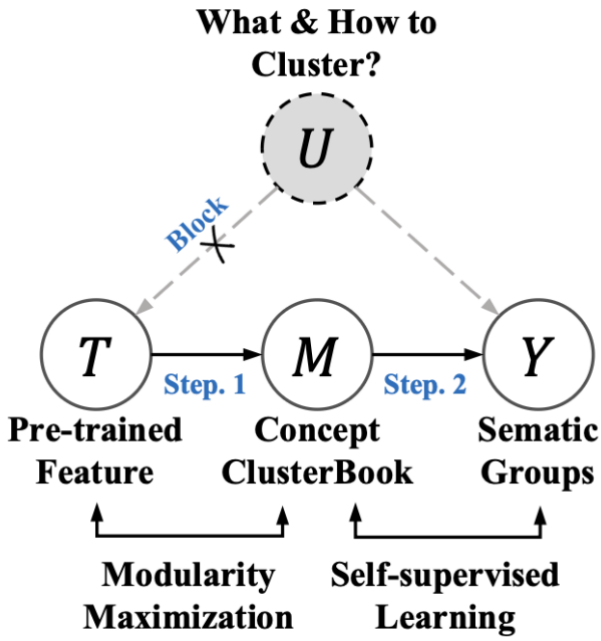


Figure 18: Causal diagram of CAUSE. The method decomposes unsupervised semantic segmentation into two steps: (1) modularity maximization to derive a concept clusterbook M from pre-trained features T, and (2) self-supervised learning to align M with semantic groups Y. The latent variable U denotes the unknown clustering objective, which is blocked by CAUSE's two-step strategy.

In CAUSE, segmentation can be viewed as the process of deriving semantically coherent groups Y from pre-trained features T. However, because the clustering objective is not explicitly defined, an indeterminate factor U (representing uncertainty about what and how to cluster) may intervene, introducing confounding effects during pixel-level grouping. This can be interpreted as a backdoor path ($T \leftarrow U \rightarrow Y$) that disrupts the intended segmentation process. The key insight of CAUSE is to block this confounding effect by introducing an intermediate representation M, a subdivided concept clusterbook with discretized indices. This mediator explicitly links T to Y in a structured way, serving both as a representation of fine-grained concept prototypes and as a supervisory signal for clustering. Intuitively, the construction of M amounts to pre-defining as many inherent semantic concepts as possible, across multiple levels of granularity, and then using them as anchors to guide segmentation. Once M is established, the model can train a segmentation head that consolidates these concept prototypes into broader categories. The discretized indices in M provide a natural mechanism for identifying positive and negative samples for any given anchor, which enables concept-wise self-supervised learning. From a causal perspective, this procedure can be interpreted

as a frontdoor adjustment, since the mediator M effectively blocks the backdoor confounding introduced by U, ensuring a valid causal path ($T \rightarrow M \rightarrow Y$).

When revisiting the above example, we can instantiate a mediator M as accumulation of tar in lungs, which only affects lung cancer Y from smoking T. We then average the probable effect between tar M and lung cancer Y across all of the participants' population T' on smoking.

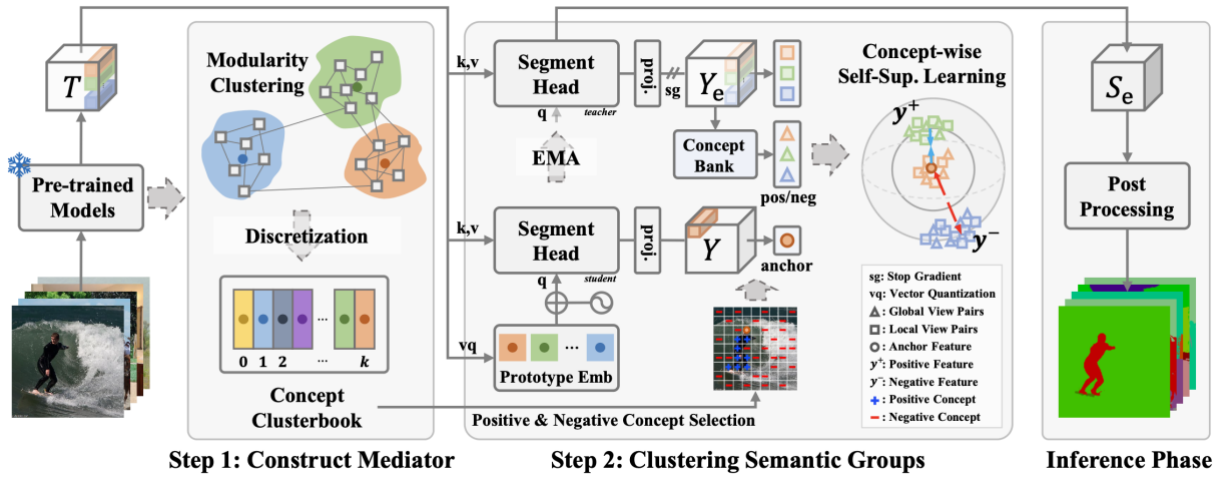


Figure 19: overall pipeline of CAUSE.

The overall workflow of CAUSE is illustrated in Figure 15. Starting from pre-trained features T, extracted from a frozen backbone, the architecture proceeds in two main stages. In the first stage, these continuous embeddings are discretized into a concept clusterbook M using modularity-based clustering, producing a mediator that bridges raw features and semantic groups. In the second stage, concept-wise self-supervised learning is applied: embeddings are aligned with their assigned prototypes, positive and negative concept relations are exploited, and a segmentation head consolidates fine-grained prototypes into higher-level semantic categories. During inference, the trained segmentation head produces semantic maps that are subsequently refined by post-processing. In the following sections, we examine these two stages in detail, first the construction of the mediator M, and then the clustering of semantic groups, before presenting the modifications required to adapt this framework to the biomedical MRI domain.

3.1.1 Step one: Maximizing Modularity

In the context of CAUSE, the starting point is a set of high-dimensional feature embeddings T, extracted from a pre-trained backbone such as DINO. Unlike raw pixel intensities, which encode only grayscale or contrast values, these embeddings represent each pixel or patch as a vector in a high-dimensional semantic space. The backbone thus serves as a powerful feature extractor, transforming raw data into representations that capture both low-level image patterns (e.g., edges, textures) and higher-level structures (e.g., tissues or pathological regions in an MRI setting). This transformation is crucial because it organizes the data into a structured and discriminative space, where similar patterns are embedded closer together while different structures are separated. Directly clustering these embeddings, however, would risk producing unstable or spurious partitions due to noise and confounding effects. To overcome this limitation, CAUSE introduces an intermediate discretized representation, called the mediator *M* or *clusterbook*. The mediator acts as a bridge between the continuous representations in T and the semantic groups Y, forcing embeddings into discrete prototypes that can later be consolidated into semantic categories. While reconstruction-based vector quantization is well-suited for generative modeling, dense prediction requires a more structured clustering strategy that accounts for pixel locality and connectivity. For this reason, CAUSE employs modularity maximization, a graph-based clustering method widely used in community detection. The idea is to model embeddings as nodes in a graph, with edges weighted by similarity (e.g., cosine similarity). The objective is to partition the graph such that nodes within a cluster are more densely connected to each other than would be expected at random.

The overall process, become the following:

1. **Start from pre-trained features T.**

   Each patch embedding is passed through a frozen pre-trained backbone (e.g., DINO in the original implementation). This gives high-dimensional feature embeddings for every patch. These embeddings are rich in semantics but not organized into categories yet.

2. **Apply clustering with modularity maximization.**

   Instead of directly clustering pixels into final segmentation labels (which is unstable), CAUSE first clusters them into many fine-grained concept groups.

   - The method builds a graph where each node is a patch embedding, and edges connect similar embeddings (based on cosine similarity).
   - Modularity maximization partitions this graph into clusters such that the density of connections within clusters is maximized, and across clusters minimized.
   - This produces a set of prototype concepts, each corresponding to a discrete cluster index in *M*.

3. **Result: the mediator *M*.**

   The output of this step is a mapping between each patch embedding and its cluster index (prototype ID). Importantly:

   - M is more fine-grained than the final semantic categories (e.g., multiple clusters may later correspond to "tumor" vs. "healthy tissue").
   - M serves as a causal mediator: it represents possible inherent concepts present in the data, and will later be consolidated into broader semantic classes during step 2.

The following formulation represents maximizing a measure of modularity *H* acquire the discretized concept fractions from pre-trained features *T*:

$$\max_{M} \mathcal{H} = \frac{1}{2e}\mathrm{Tr}\left(\mathcal{C}(T,M)^{\mathrm{T}}\left[\mathcal{A}(T) - \frac{dd^{\mathrm{T}}}{2e}\right]\mathcal{C}(T,M)\right) \in \mathbb{R}$$

where *C(T, M)* denotes cluster assignment matrix such that *max(0,cos(T,M))* between all *hw* patch feature points in pre-trained features *T* and all *k* concept prototypes in *M*. The cluster assignment matrix implies how each patch feature point is close to concept prototypes. In addition, *A(T)* indicates the affinity matrix of *T* such that $A_{ij} = \max\left(0,\ \cos\left(t_i, t_j\right)\right)$ between the two patch

feature points t_i, t_j in *T*, which represents the intensity of connections among vertices. Note that, the degree vector *d* describes the number of the connected edges in its affinity *A*, and *e* denotes the total number of the edges. By jointly considering cluster assignments *C(T,M)* and affinity matrix *A(T)* at once, in brief, maximizing modularity *H* constructs the discretized concept clusterbook *M* taking into account the patch-wise locality and connectivity in pre-trained representation *T*. The procedure for constructing the mediator *M* through modularity maximization is summarized in **Algorithm 1**

Algorithm 1 (STEP 1) Maximizing Modularity for Constructing Concept Clusterbook M

Require: Image Samples $X \sim$ Data, Pre-trained Model f, Concept Fractions $M \in \mathbb{R}^{k \times c}$

Initialize M
for $X \sim$ Data **do**
  $T \in \mathbb{R}^{hw \times c} \leftarrow f(X)$ ▷ Pre-trained Model Representation
  $\mathcal{A} \leftarrow \max(0, \cos(T, T)) \in \mathbb{R}^{hw \times hw}$ ▷ Affinity matrix
  $d, e \leftarrow \mathcal{A}$ ▷ Degree Matrix and Number of Total edges
  $\mathcal{C} \leftarrow \max(0, \cos(T, M)) \in \mathbb{R}^{hw \times k}$ ▷ Cluster Assignment Matrix
  $\mathcal{H} \leftarrow \frac{1}{2e}\mathrm{Tr}\left(\tanh\left(\frac{\mathcal{C}\mathcal{C}^T}{\tau}\right)\left[\mathcal{A} - \frac{dd^T}{2e}\right]\right)$ ▷ Maximizing Modularity ($\tau = 0.1$)
  $M \leftarrow \mathrm{Increase}(\mathcal{H})$ ▷ Updating Concept ClusterBook (lr: 0.001)
end for

As a result, each patch in the image is represented not by its original embedding, but by the index of the prototype to which it is assigned. Importantly, this clusterbook is designed to capture concepts at multiple granularities. Fine-grained prototypes correspond to localized textures or small-scale structures, while coarser prototypes encode broader semantic categories. By subdividing the feature space in this way, the mediator *M* ensures that subsequent learning stages are grounded in a set of predefined and consistent semantic units. From a causal perspective, the introduction of *M* acts as a blocking mechanism against confounding effects arising from the indeterminate nature of unsupervised clustering. Without such a mediator, clustering would be ill-posed, as multiple groupings of the same features could emerge without any principled way to resolve them. By enforcing discretization into structured prototypes, CAUSE establishes a valid causal pathway from the features *T* to the semantic groups *Y* via the mediator M, satisfying the conditions for frontdoor adjustment.

3.1.2. Step two: Enhancing Likelihood of Semantic Groups

Once the mediator M has been established, CAUSE proceeds to the second stage, where concept-wise self-supervised learning is employed to refine and consolidate prototype assignments into meaningful semantic clusters. The central idea of this step is to embed segmentation features *Y* that align with the concept prototypes from the pre-trained features *T*, achieved by training a task-specific prediction head *S*. This head receives patch-level features *T* and performs cross-attention with the concept prototypes stored in *M*, producing concept-matched embeddings. In practice, each pixel embedding is associated with a prototype index derived from *M*, and the segmentation head consolidates these fine-grained prototypes into broader semantic categories.

The training process is framed as a contrastive learning task. For a given anchor embedding, other embeddings assigned to the same prototype act as positive samples, while embeddings tied to different prototypes serve as negative samples. This creates supervisory signals without requiring manual annotations. Technically, this is expressed through a Noise-Contrastive Estimation (NCE) loss, where the probability of associating a pixel feature with its positive prototype is maximized, while simultaneously minimizing its similarity with negatives.

$$p(y \mid t', m) = \mathop{\mathbb{E}}_{y^+} \left[\frac{\exp(\cos(y, y^+)/\tau)}{\exp(\cos(y, y^+)/\tau) + \sum_{y^-} \exp(\cos(y, y^-)/\tau)} \right]$$

Where y, y^+, y^- denote anchor, positive, and negative features and τ indicates temperature terms.

To improve robustness, CAUSE introduces a positive and negative concept selection mechanism: based on a pre-computed similarity matrix between prototypes, only sufficiently close prototypes (above a positive threshold) are retained as positives, while sufficiently dissimilar ones (below a negative threshold) are treated as negatives. This selective relaxation stabilizes training by avoiding ambiguous cases.

To further enhance learning, CAUSE incorporates a concept bank strategy. Unlike typical instance-level contrastive learning, where comparisons are limited to a single batch, the concept bank accumulates positive and negative concept features across multiple mini-batches, thus broadening the context available for contrastive learning. This mechanism ensures that each prototype is

continuously refined against a diverse set of examples, effectively simulating global-level supervision. The result is a stronger clustering signal and more stable feature separation.

From a causal perspective, the use of M ensures that the predictive mapping follows the valid frontdoor pathway $(T \rightarrow M \rightarrow Y)$ while blocking spurious backdoor paths ($T \leftarrow U \rightarrow Y$) that could bias unsupervised clustering. This combination of contrastive learning, selective positive/negative sampling, and accumulated context via the concept bank grounds the method in both solid technical design and causal inference theory. In biomedical imaging, this means that prototypes capturing subtle textural variations in tissues (e.g., different subtypes of white matter) can be robustly grouped into meaningful anatomical categories such as gray matter or white matter, with improved stability and interpretability.

3.1.3 Evidence and Discussion

After introducing the conceptual foundation of CAUSE, it is essential to review its empirical validation and the insights derived from the original work. This section synthesizes the implementation details, dataset choices, evaluation metrics, ablation studies, and generalization analyses that establish CAUSE as a state-of-the-art framework in unsupervised semantic segmentation.

CAUSE adopts DINO as the baseline encoder, with its weights frozen during training. Depending on the Vision Transformer (ViT) size, the feature dimension ccc of the encoder is either 384 (ViT-Small) or 768 (ViT-Base). The concept clusterbook M is set to 2048 prototypes to ensure sufficient coverage of semantic concepts extracted from pre-trained features. During self-supervised learning, the concept bank accumulates 100 feature samples per prototype, and the segmentation head has an output dimension of 90. Two variants of segmentation heads are considered: CAUSE-MLP, consisting of simple MLP layers, and CAUSE-TR, which employs a single-layer transformer.

CAUSE is primarily benchmarked on three widely used semantic segmentation datasets: COCO-Stuff, Cityscapes, and Pascal VOC. COCO-Stuff is derived from MS-COCO 2017 and provides dense pixel annotations for common “stuff” and “thing” categories. Cityscapes focuses on urban scene parsing with fine-grained pixel annotations, while Pascal VOC serves as an object-centric benchmark. Performance is assessed using mean Intersection-over-Union (mIoU) and pixel accuracy (pAcc). Across datasets, CAUSE consistently outperforms recent unsupervised

frameworks, including STEGO, HP [65], and the others previously cited. Beyond numerical superiority, qualitative evaluations reveal that CAUSE is particularly effective in consolidating fine-grained prototypes (e.g., head, torso, hand, leg) into semantically coherent higher-level categories (e.g., person), demonstrating its capacity to balance detail and abstraction.

An important advantage of CAUSE lies in its adaptability across different self-supervised encoders. Substituting DINO with alternatives such as DINOv2, iBOT, and MAE confirms the robustness of the framework. Furthermore, scaling the number of clusters to match MS-COCO 2017's 171 categories (80 objects + 91 "stuff") shows that CAUSE maintains competitive performance, even with significantly larger label spaces. This robustness highlights the causal grounding of the method, which reduces reliance on specific architectural choices or datasets.

To understand the contribution of individual components, extensive ablation studies are conducted. Six key factors are examined:

1. Positive (ϕ+\phi^+ϕ+) and negative (ϕ−\phi^-ϕ−) relaxation parameters,
2. The number of prototypes *k* in the clusterbook *M*,
3. The role of the concept bank,
4. The effect of using fully-connected CRFs for refinement, and
5. Different discretization strategies for M.

(a) Experimental results on COCO-Stuff.

Method ($\mathbb{C} = 27$)	Backbone	mIoU	pAcc
IIC (Ji et al., 2019)	ResNet18	6.7	21.8
PiCIE (Cho et al., 2021)	ResNet18	14.4	50.0
SegDiscover (Huang et al., 2022)	ResNet50	14.3	40.1
SlotCon (Wen et al., 2022)	ResNet50	18.3	42.4
HSG (Ke et al., 2022)	ResNet50	23.8	57.6
ReCo+ (Shin et al., 2022)	DeiT-B/8	32.6	54.1
DINO (Caron et al., 2021)	ViT-S/16	8.0	22.0
+ STEGO (Hamilton et al., 2022)	ViT-S/16	23.7	52.5
+ HP (Seong et al., 2023)	ViT-S/16	24.3	54.5
+ **CAUSE-MLP**	ViT-S/16	25.9	66.3
+ **CAUSE-TR**	ViT-S/16	**33.1**	**70.4**
DINO (Caron et al., 2021)	ViT-S/8	11.3	28.7
+ ACSeg (Li et al., 2023)	ViT-S/8	16.4	-
+ TranFGU (Yin et al., 2022)	ViT-S/8	17.5	52.7
+ STEGO (Hamilton et al., 2022)	ViT-S/8	24.5	48.3
+ HP (Seong et al., 2023)	ViT-S/8	24.6	57.2
+ **CAUSE-MLP**	ViT-S/8	27.9	66.8
+ **CAUSE-TR**	ViT-S/8	**32.4**	**69.6**
DINO (Caron et al., 2021)	ViT-B/8	13.0	42.4
+ DINOSAUR (Seitzer et al., 2023)	ViT-B/8	24.0	-
+ STEGO (Hamilton et al., 2022)	ViT-B/8	28.2	56.9
+ **CAUSE-MLP**	ViT-B/8	34.3	72.8
+ **CAUSE-TR**	ViT-B/8	**41.9**	**74.9**

(b) Experimental results on Cityscapes.

Method ($\mathbb{C} = 27$)	Backbone	mIoU	pAcc
IIC (Ji et al., 2019)	ResNet18	6.4	47.9
PiCIE (Cho et al., 2021)	ResNet18	10.3	43.0
SegSort (Hwang et al., 2019)	ResNet101	12.3	65.5
SegDiscover (Huang et al., 2022)	ResNet50	24.6	81.9
HSG (Ke et al., 2022)	ResNet50	32.5	86.0
ReCo+ (Shin et al., 2022)	DeiT-B/8	24.2	83.7
DINO (Caron et al., 2021)	ViT-S/8	10.9	34.5
+ TransFGU (Yin et al., 2022)	ViT-S/8	16.8	77.9
+ HP (Seong et al., 2023)	ViT-S/8	18.4	80.1
+ **CAUSE-MLP**	ViT-S/8	21.7	87.7
+ **CAUSE-TR**	ViT-S/8	**24.6**	**89.4**
DINO (Caron et al., 2021)	ViT-B/8	15.2	52.6
+ STEGO (Hamilton et al., 2022)	ViT-B/8	21.0	73.2
+ HP (Seong et al., 2023)	ViT-B/8	18.4	79.5
+ **CAUSE-MLP**	ViT-B/8	25.7	90.3
+ **CAUSE-TR**	ViT-B/8	**28.0**	**90.8**

(c) Self-supervised methods with CAUSE-TR.

Dataset	Self-Supervised Methods	Backbone	mIoU	pAcc
COCO-Stuff			45.3	78.0
Cityscapes	DINOv2 (Oquab et al., 2023)	ViT-B/14	29.9	89.8
Pascal VOC			53.2	91.5
COCO-Stuff			39.5	73.8
Cityscapes	iBOT (Zhou et al., 2022)	ViT-B/16	23.0	89.1
Pascal VOC			53.4	89.6
COCO-Stuff			34.1	72.1
Cityscapes	MSN (Assran et al., 2022)	ViT-S/16	21.2	89.1
Pascal VOC			30.2	84.2
COCO-Stuff			21.5	59.1
Cityscapes	MAE (He et al., 2022)	ViT-B/16	12.5	82.0
Pascal VOC			25.8	83.7

(d) Experimental results on Pascal VOC 2012.

Method ($\mathbb{C} = 21$)	Backbone	mIoU
IIC (Ji et al., 2019)	ResNet18	9.8
SegSort (Hwang et al., 2019)	ResNet101	11.7
DenseCL (Wang et al., 2021b)	ResNet50	35.1
HSG (Ke et al., 2022)	ResNet50	41.9
MaskContrast (Van Gansbeke et al., 2021)	ResNet50	35.0
MaskDistill (Van Gansbeke et al., 2022)	ResNet50	48.9
DINO (Caron et al., 2021)	ViT-S/8	-
+TransFGU (Yin et al., 2022)	ViT-S/8	37.2
+ACSeg (Li et al., 2023)	ViT-S/8	47.1
+CAUSE-MLP	ViT-S/8	46.0
+CAUSE-TR	ViT-S/8	**50.0**
DINO (Caron et al., 2021)	ViT-B/8	-
+DeepSpectral (Melas-Kyriazi et al., 2022)	ViT-B/8	37.2
+DINOSAUR (Seitzer et al., 2023)	ViT-B/8	37.2
+Leopart (Ziegler & Asano, 2022)	ViT-B/8	41.7
+COMUS (Zadaianchuk et al., 2023)	ViT-B/8	50.0
+CAUSE-MLP	ViT-B/8	47.9
+CAUSE-TR	ViT-B/8	**53.3**

Figure 20: Quantitative comparison on COCO-Stuff, Cityscapes, and Pascal VOC. CAUSE consistently improves performance across different backbones and datasets, with CAUSE-TR achieving state-of-the-art results.

Results indicate that relaxation parameters are crucial for balancing semantic coherence and separation, while the number of prototypes saturates around 2048, ensuring sufficient semantic granularity without redundancy. The concept bank, which accumulates features across batches, proves vital for stabilizing self-supervised training and improving coverage of rare concepts.

3.2 Domain Adaptation

One of the most meaningful modifications introduced in this thesis concerns the adaptation of the CAUSE pipeline to handle 2D and 3D MRI data stored in NIfTI format. While the original CAUSE framework was designed with natural images in mind, medical imaging requires an entirely different preprocessing pipeline due to its volumetric nature, modality heterogeneity, and strict

requirements for spatial consistency. To address this, I developed a dedicated data loader, CroppedMRIDataset, which extends CAUSE to process MRI volumes and generate tensors suitable for weakly supervised segmentation in both 2D and 3D settings.

The implementation leverages the MONAI framework (Medical Open Network for AI), an open-source PyTorch-based library tailored for healthcare imaging. MONAI provides specialized data handling and preprocessing utilities that standardize operations such as voxel resampling, orientation alignment, and multi-channel support, thereby ensuring that MRI volumes from heterogeneous sources can be integrated into a unified pipeline.

The main preprocessing steps included in the adapted pipeline are as follows:

- NIfTI Loading and Channel-First Formatting. Both images and labels are loaded from NIfTI files and converted into tensors. Images are stored in *[C, D, H, W]* format, while labels are represented as *[D, H, W]* or *[1, D, H, W]*. This ensures compatibility with 3D convolutional or transformer-based backbones.

- Voxel Spacing Normalization. All volumes are resampled to a target isotropic voxel size (sx, sy, sz). For continuous intensity images, bilinear interpolation is used, while categorical labels are resampled with nearest-neighbor interpolation to preserve class integrity. This step standardizes the spatial resolution across different scanners and acquisition protocols.

- Orientation Normalization. Volumes are reoriented into the RAS coordinate system (Right–Anterior–Superior) to ensure consistent axis ordering across patients and institutions. Without this step, the same anatomical structure might appear along different axes depending on scanner configuration, which would introduce spurious variability into the training process.

- Intensity Normalization. Raw MRI intensities are highly variable across scanners and acquisition settings. To mitigate this, voxel intensities are clipped and rescaled into the range *[0, 1]* using a predefined window. This improves training stability and facilitates cross-subject comparability.

- Foreground Cropping. Volumes are cropped to the smallest bounding box that encloses all non-zero voxels, thereby removing empty background regions. The same crop is applied to both image and label. This not only reduces computational overhead but also improves focus on clinically relevant regions, which is critical for tumor segmentation.

- Optional 3D Augmentation. To increase robustness, optional volumetric augmentations are applied consistently to both image and label, including random flips along axes, 90° rotations, and Gaussian noise injection (image only). These augmentations expand the effective dataset size and expose the model to plausible anatomical variability.

- Multi-Modal Support. The loader is capable of handling multi-channel MRI stored in a single NIfTI file (e.g., FLAIR, T1, T1ce, and T2), automatically preserving channel information in the *[C, D, H, W]* format. This is essential for BraTS, where tumor subregions are visible only under specific modalities.

The resulting preprocessed images and labels are returned as 3D tensors ready for training. To ensure reproducibility and flexibility, a default configuration file was implemented, allowing customization of voxel spacing, intensity windowing, and augmentation parameters.

A particularly important component of this preprocessing pipeline is also the cropping strategy. Beyond reducing computational costs, cropping enables the use of multi-scale crops, which serve as a self-supervised signal. By presenting the network with both global crops (covering the whole anatomical context) and local crops (focused on finer structures), the model learns to integrate information across scales. This is especially beneficial in 3D MRI, where accurate segmentation requires simultaneously capturing global context (the position and relation of a tumor within the brain) and fine local detail (precise delineation of tumor boundaries). In the context of CAUSE, this design choice reinforces the alignment between self-supervised pretraining principles and the demands of biomedical segmentation.

Another crucial element of the CAUSE framework is the choice of backbone, since it provides the high-dimensional feature representations on which the entire causal pipeline relies. In the first stage of this work, the selected backbone was DINOv2. DINOv2 is trained on the LVD-142M dataset, a carefully curated collection of approximately 142 million natural images. Of these, roughly 100 million were selected based on their similarity to ImageNet21K samples, while the remainder were

retrieved by similarity to well-known datasets such as Caltech-101, ADE20K, and Google Landmarks v2. The resulting dataset is therefore highly diverse, and this characteristic makes DINOv2 particularly interesting for medical imaging research: it allows to rigorously test the cross-domain generalization capabilities of large-scale self-supervised models, i.e., their ability to transfer learned visual priors from natural images to highly distinct biomedical data distributions.

The decision to start with DINOv2 was strongly motivated by recent work evaluating its performance in the medical domain. In particular, a comprehensive experimental study [66] systematically assessed the backbone across multiple radiology modalities (X-ray, CT, MRI) and tasks, including both classification and organ segmentation, under different data regimes ranging from few-shot to large-scale training. Importantly, the study confirmed that DINOv2 performs remarkably well "out of the box": even when the backbone is kept frozen and only lightweight decoders are trained, DINOv2 outperformed or matched other foundation models such as CLIP, OpenCLIP, MAE, and MSN across numerous benchmarks. These results highlight two key aspects that directly motivated the following implementation choice. First, the self-distillation training strategy of DINOv2 yields feature spaces that are semantically well-structured. Second, the model's ability to provide competitive performance with frozen weights is especially appealing in biomedical imaging, where annotated data are scarce, and computational budgets may limit the feasibility of large-scale fine-tuning.

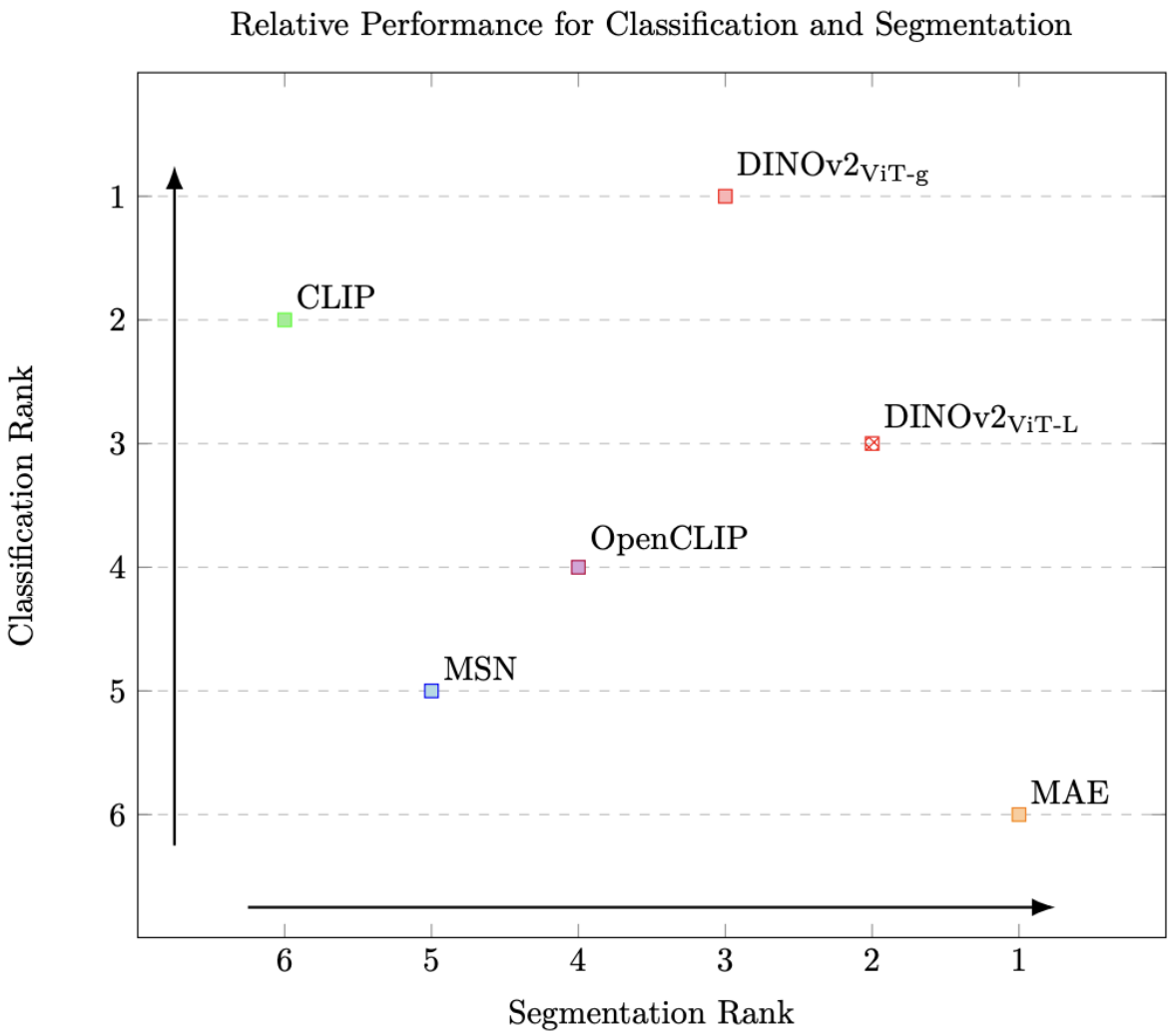


Figure 23: Cross-task generalizability of DINOv2 compared to other models. The horizontal axis shows the relative performance or ranking of each of the models on segmentation tasks, while the vertical axis does the same for classification tasks. The ranking is calculated as the average ranking across all segmentation (5 Datasets) or classification tasks (4 Datasets), where rank 1 means the model performs the best relative to other models.

In this work, we rely on the official pre-trained weights of DINOv2, integrating them as a frozen feature extractor within CAUSE to test how far a general-purpose model trained entirely on natural images can support unsupervised clustering and causal reasoning in the domain of MRI semantic segmentation. It should be noted that DINOv3, the successor to DINOv2, was released only during the final stages of writing this thesis and was therefore not included in the experimental pipeline, although it is discussed in the Background chapter and proposed as a natural extension for future work.

While DINOv2 was an excellent choice to probe the generalization capacity of natural image foundation models, its limitations in the biomedical setting became clear: although the learned features are rich, they are not specifically tailored to capture the volumetric and multi-modal characteristics of MRI data. For this reason, a second backbone was explored, this time explicitly pre-trained on medical imaging data. After surveying the limited set of available medical-native self-supervised transformers, the chosen model was Swin-UNETR [67], developed within the MONAI research framework.

Swin-UNETR integrates a 3D Swin Transformer encoder with a U-Net style decoder, combining the strengths of hierarchical self-attention and skip connections for fine-grained boundary preservation. Unlike DINOv2, which was trained on natural images, Swin-UNETR was pre-trained self-supervised on over 5,000 CT volumes using a combination of proxy tasks specifically designed for medical data: masked volume inpainting (predicting missing sub-volumes from context), 3D rotation prediction (learning orientation-invariant features), and contrastive learning across slices and augmentations (to enforce volumetric consistency). This training strategy imbued the model with inductive biases aligned with the properties of medical images, enabling it to better capture both fine anatomical details and long-range 3D dependencies across volumetric scans.

From an architectural perspective, the 3D Swin Transformer encoder processes volumes hierarchically through shifted window attention, efficiently capturing both local and global context. These representations are then passed to a U-shaped decoder with skip connections, which reintroduce high-resolution spatial information to refine boundary delineations. Crucially, Swin-UNETR can natively ingest multi-modal MRI volumes (e.g., FLAIR, T1, T1ce, T2), treating them as channels in a 3D tensor, and preserve their complementary information throughout the encoder–decoder pipeline. This ability to directly handle the volumetric and multi-modal nature of MRI data makes Swin-UNETR particularly well suited for integration into CAUSE, where the mediator M relies on the richness and stability of the input features. In practice, this meant that the fine-grained prototypes produced in Step 1 were better aligned with anatomical structures, while the segmentation head in Step 2 could leverage volumetric coherence to improve semantic consolidation.

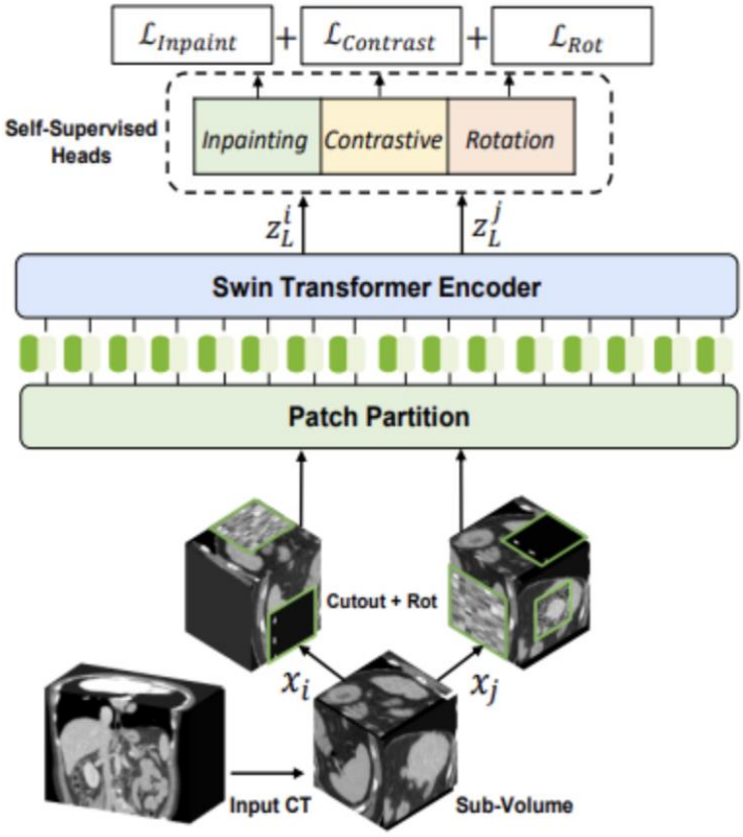


Figure 24: encoder of Swin-UNETR

The exploration of Swin-UNETR therefore represents a crucial adaptation beyond the original CAUSE framework. While the DINOv2 experiments served to test how far general-purpose natural image representations could be stretched into the medical domain, Swin-UNETR anchors the pipeline within the biomedical field itself, aligning both the pretraining data and the architectural design with the unique challenges of MRI segmentation. By comparing these two backbones—one generalist and one specialist—this thesis provides insights into the trade-offs between cross-domain transfer and domain-specific inductive bias in the context of weakly supervised biomedical image segmentation.

3.3 Main Contributions

After having detailed the adaptations made to the CAUSE framework for its application in the biomedical domain, it is important to situate this contribution within the broader research context of unsupervised semantic segmentation in medical imaging. Unsupervised biomedical

segmentation remains a relatively underexplored area compared to supervised and semi-supervised approaches, which continue to dominate the field. Most state-of-the-art methods in medical imaging either rely on large annotated datasets, which are costly and time-consuming to obtain, or adopt weak supervision strategies such as scribbles, bounding boxes, or image-level labels to reduce annotation requirements. While these approaches have shown promise, they still depend on some form of human annotation, leaving open the question of how to achieve robust segmentation without labels at all. By contrast, unsupervised methods developed in the general computer vision community—often centered around self-supervised learning and clustering—have rarely been systematically transferred to the biomedical domain, in part due to the unique characteristics of medical images (multi-modality, low contrast, domain variability, and clinical relevance of fine boundaries).

This thesis contributes to filling this gap by adapting the CAUSE framework. The novelty lies in two complementary aspects: (i) bringing causal inference into unsupervised biomedical segmentation, where causal reasoning provides a principled way to mitigate confounding factors and stabilize clustering-based predictions, and (ii) integrating this framework with self-supervised Vision Transformer backbones, which are capable of learning rich and transferable representations without labels. The implementation is evaluated on two benchmark datasets, the LGG dataset (2D binary tumor segmentation) and the BraTS 2016–2017 dataset (3D multi-class tumor subregion segmentation), which allows us to investigate both simpler and more complex scenarios.

Compared to existing literature, this work therefore goes beyond the predominant reliance on supervised and semi-supervised pipelines by exploring a purely unsupervised strategy tailored for MRI segmentation, demonstrating both the opportunities and the limitations of such an approach. It shows that while general-purpose self-supervised Transformers such as DINOv2 can provide strong baselines, domain-specific architectures like Swin-UNETR significantly improve results, highlighting the importance of domain adaptation. In doing so, this thesis extends the frontier of research by providing empirical evidence and methodological insights into how causality and self-supervision can be jointly leveraged for one of the most challenging but impactful tasks in biomedical computer vision.

In the next chapter, we will present the datasets employed and the results obtained with the proposed method, analyzing both its strengths and the remaining open challenges.

4. Data and Results

Building on the methodological adaptations presented in the previous chapter, namely, the integration of CAUSE with MRI data through tailored preprocessing, the use of different backbones, and the careful adjustment of training parameters, this section presents the datasets used as well as the experimental outcomes obtained.

4.1 MRI Images and Datasets

Magnetic Resonance Imaging (MRI) is one of the most widely used imaging modalities in clinical practice, providing non-invasive, high-resolution insights into soft tissues, organs, and pathological structures. MRI relies on the interaction between strong magnetic fields and the intrinsic magnetic properties of atomic nuclei, particularly protons within water molecules, which are abundant in biological tissues.

The decision to focus this work on MRI data is motivated by both its clinical and technical significance. Clinically, MRI is one of the most widely used imaging modalities in neurology and oncology: its central role in the diagnosis, monitoring, and treatment planning of conditions such as brain tumors makes it a critical application domain for semantic segmentation. From a technical perspective, MRI presents a set of challenges that are particularly well aligned with the research question addressed in this thesis. Unlike natural images, which are typically two-dimensional, MRI scans capture anatomy in three dimensions, often across multiple modalities or sequences (e.g., T1, T2, FLAIR). Each MRI volume can consist of hundreds of slices, resulting in a 3D voxel grid rather than a 2D pixel matrix. This volumetric nature is crucial for preserving anatomical continuity but dramatically increases computational and memory requirements. The spatial resolution along the axial plane (x–y) may differ from the inter-slice resolution (z-axis), producing anisotropic voxels that complicate uniform processing and model design. Moreover, MRIs often exhibit low signal-to-noise ratios and poor contrast between adjacent tissues, particularly in brain imaging, where gray matter and white matter boundaries can be subtle. Pathological regions, such as tumors, can blend into surrounding tissue. Different MRI sequences highlight different tissue properties: T1-weighted scans provide structural detail, T2-weighted emphasize fluid and edema, while FLAIR suppresses

fluid signals to enhance lesion visibility. A clinically meaningful segmentation task often requires integrating information across modalities. This multimodal dependency introduces complexity absent in single-channel natural images and necessitates architectural designs capable of fusing complementary signals. Overall, these peculiarities demand significant modifications compared to the original CAUSE pipeline.

To evaluate the proposed framework, I deliberately selected two public brain MRI datasets that differ substantially in size, complexity, and annotation granularity, thereby enabling a progressive validation strategy.

The choice to work with two datasets is motivated by the need to test the framework across diverse conditions that reflect real-world biomedical imaging scenarios: from small, binary, slice-based datasets, to large, multi-institutional, volumetric datasets with multi-class annotations. This dual evaluation not only helps to assess the adaptability of the method under varying constraints but also mirrors the clinical diversity of data encountered in practice. I began my experimentation with the LGG MRI Segmentation dataset available on Kaggle, a collection of scans from patients with lower-grade gliomas (LGG). This dataset is composed of T1-weighted brain MRI slices with corresponding binary tumor masks provided by clinical experts. It contains approximately 110 cases, from these 3D volumes, axial slices were extracted, resulting in over 3,000 2D images with corresponding tumor annotations. Unlike volumetric datasets, LGG is structured as a set of 2D axial slices, which considerably reduces the computational burden and simplifies the design of early experiments. The dataset is therefore organized as a collection of paired 2D slices and masks, making it computationally lightweight and suitable for rapid experimentation.

Its relatively small scale made it an accessible starting point for prototyping the adaptation of CAUSE, particularly since the binary segmentation task (tumor versus background) allowed me to focus on validating the feasibility of the method under conditions of annotation scarcity. Nevertheless, the dataset is far from trivial: gliomas in this collection exhibit wide variability in shape, size, and intensity profiles, often with low-contrast and ambiguous boundaries against healthy tissue. These characteristics make segmentation challenging even for expert radiologists, highlighting the very limitations of supervised methods that the proposed weakly/self-supervised approach seeks to address.

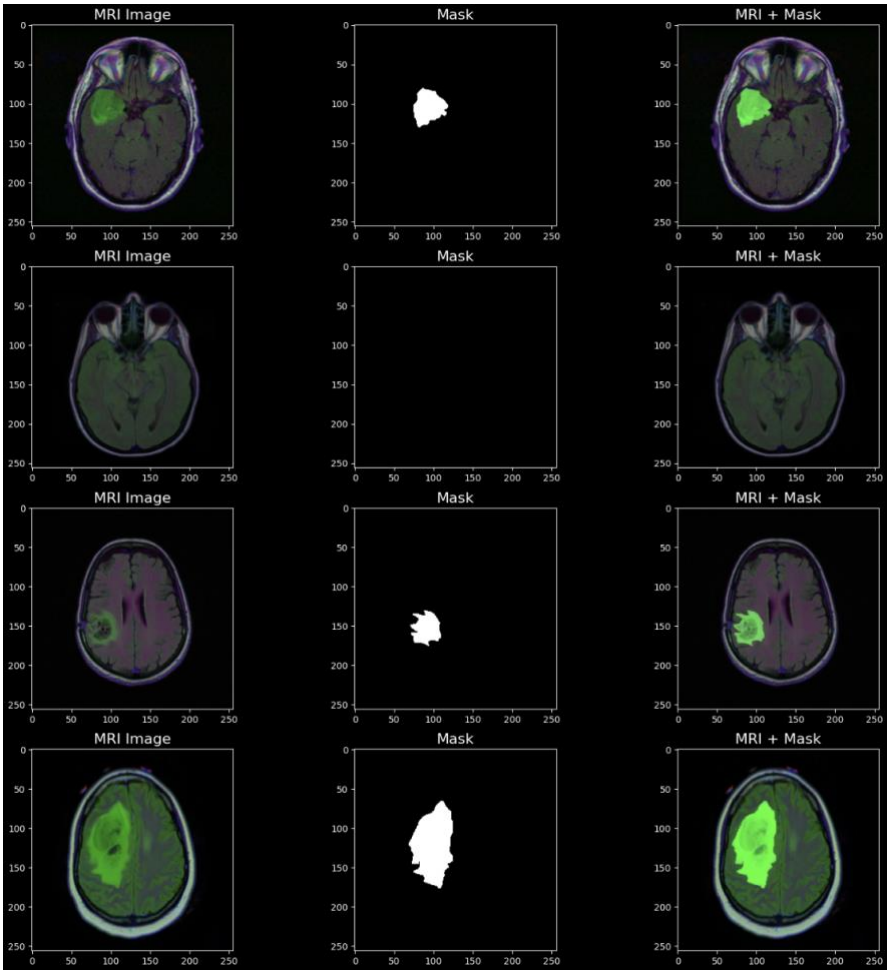


Figure 21: example MRI slices from the LGG Kaggle dataset. Each row shows the original brain MRI (left), the ground-truth tumor mask (center), and the overlay of mask on the MRI (right), highlighting tumor regions in patients with lower-grade glioma.

After these preliminary experiments, I extended the evaluation to the Brain Tumor Segmentation (BraTS) challenges of 2016 and 2017, distributed through the Medical Segmentation Decathlon (Task 1: Brain Tumor). The BraTS datasets represent one of the most widely recognized and demanding benchmarks in medical image segmentation. Unlike LGG, BraTS provides full 3D volumetric MRI scans acquired from multiple institutions and scanners, introducing significant domain variability. In total, the dataset includes 750 multi-parametric MRI scans from patients diagnosed with either glioblastoma or lower-grade glioma, collected across 19 different institutions. Each case consists of four MRI modalities: native T1-weighted, contrast-enhanced T1 (T1-Gd), T2-weighted, and FLAIR. These sequences provide complementary information — for example, T1-Gd highlights enhancing tumor regions, while FLAIR is effective in identifying edema. The ground-truth annotations are more detailed than those in LGG, as they delineate three clinically relevant tumor subregions: (1) the enhancing tumor, (2) the necrotic or non-enhancing core, and

(3) the edema. This structure transforms the task into a multi-class segmentation problem, considerably more complex than the binary masks provided in LGG. Furthermore, the data are fully 3D volumetric scans, with each volume resampled to isotropic 1 mm^3 voxel resolution and co-registered to the SRI24 brain atlas template. Preprocessing also included skull-stripping and manual refinements to ensure anatomical consistency.

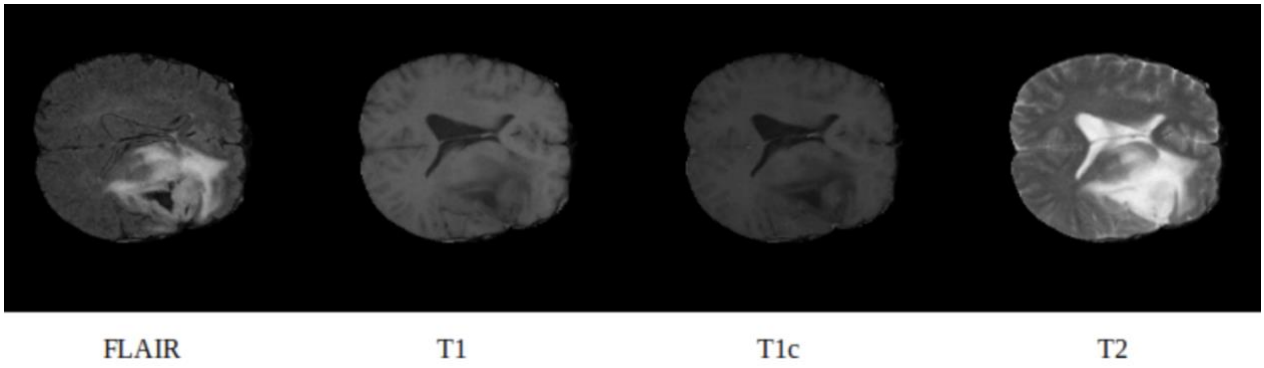


Figure 22: Example case from the BraTS dataset, showing the four MRI modalities provided for each patient: FLAIR, native T1, contrast-enhanced T1 (T1-Gd), and T2.

This multi-class segmentation task is considerably more complex than binary tumor delineation, requiring the model to simultaneously integrate multimodal information, preserve volumetric consistency across slices, and discriminate subtle textural differences within heterogeneous tumor tissue.

Taken together, the LGG and BraTS datasets provide complementary testbeds. LGG is smaller, 2D, and binary, serving as a natural starting point to explore the feasibility of data-efficient segmentation in a constrained setting where annotations are scarce and tumor boundaries are subtle. BraTS, by contrast, is larger, 3D, multimodal, and multi-class, representing a clinically realistic challenge where robust generalization and fine-grained segmentation are essential. By employing both datasets, I was able to evaluate the proposed framework along a continuum of increasing difficulty: first demonstrating its applicability in a simplified binary setting, then validating its scalability and robustness in a more complex, clinically relevant benchmark. This progression mirrors the trajectory of biomedical imaging research itself, where proof-of-concept methods are often first tested on small, constrained datasets before being deployed on large, heterogeneous, real-world collections.

4.2 Evaluation

The evaluation is structured in two parts: first, analyze the results obtained using DINOv2 as the backbone, and then report the performance achieved with Swin-UNETR. For each backbone, experiments were conducted initially on the LGG Kaggle dataset as a proof of concept on a simpler 2D and binary segmentation task, and then extended to the more challenging BraTS 2016–2017 datasets, which involve 3D multi-modal MRI scans and multi-class tumor subregion segmentation.

Across all implementations, several hyperparameters were varied to optimize performance. The feature dimension (c) of the pre-trained backbone was adjusted according to the model size, with ViT-small (c = 384) and ViT-base (c = 768) being tested. Another critical parameter was the number of clusters (k) in the concept clusterbook, which was varied from 2 to 2048 in order to encompass the widest possible range of semantic prototypes from the pre-trained features. The concept bank accumulation size (b) was fixed at 100, consistent with the original CAUSE implementation, while the segmentation head output dimension (r) was set to 90 based on dimensionality analysis. Two segmentation head variations were tested: (i) CAUSE-MLP, using simple multilayer perceptron, and (ii) CAUSE-TR, using a single-layer transformer.

To evaluate segmentation performance, three primary metrics were considered:

- **Mean Intersection over Union (mIoU):** a standard metric for semantic segmentation that measures the overlap between predicted and ground truth segmentation masks, averaged across classes.
- **Mean Average Precision (mAP):** a metric evaluating how well the model distinguishes positive predictions across thresholds, providing a broader assessment of classification and localization quality.
- **Mean Dice (DSC):** overlap metric widely used in medical imaging, defined per class as

  $$\mathrm{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}$$ *(1.0 = perfect overlap, 0.0 = no overlap)*

  I compute Dice per class (after Hungarian matching for unsupervised outputs) and report the mean across classes (and averaged over test subjects). Dice is robust under class imbalance (e.g., small lesions) and complements mIoU.

The following section presents the best performances obtained across the two datasets (LGG and BraTS 2016–2017) after fine-tuning the main hyperparameters of CAUSE with different backbones.

LGG dataset (2D, binary segmentation):

- DINOv2 + CAUSE-MLP → mIoU ≈ **43%**, mAP ≈ **51%, DSC** ≈ 0.54%
- DINOv2 + CAUSE-TR → mIoU ≈ **41%**, mAP ≈ **48%, DSC** ≈ 0.53 %
- Swin-UNETR + CAUSE-MLP → mIoU ≈ **45%**, mAP ≈ **50%, DSC** ≈ 0.63%
- Swin-UNETR + CAUSE-TR → mIoU ≈ **45%**, mAP ≈ **52%, DSC** ≈ 0.62%

BraTS 2016–2017 datasets (3D, multi-class segmentation):

- DINOv2 + CAUSE-MLP → mIoU ≈ **40%**, mAP ≈ **45%, DSC** ≈ 0.55%
- DINOv2 + CAUSE-TR → mIoU ≈ **38%**, mAP ≈ **44%, DSC** ≈ 0.53%
- Swin-UNETR + CAUSE-MLP → mIoU ≈ **46%**, mAP ≈ **50%, DSC** ≈ 0.67%
- Swin-UNETR + CAUSE-TR → mIoU ≈ **43%**, mAP ≈ **49%, DSC** ≈ 0.64%

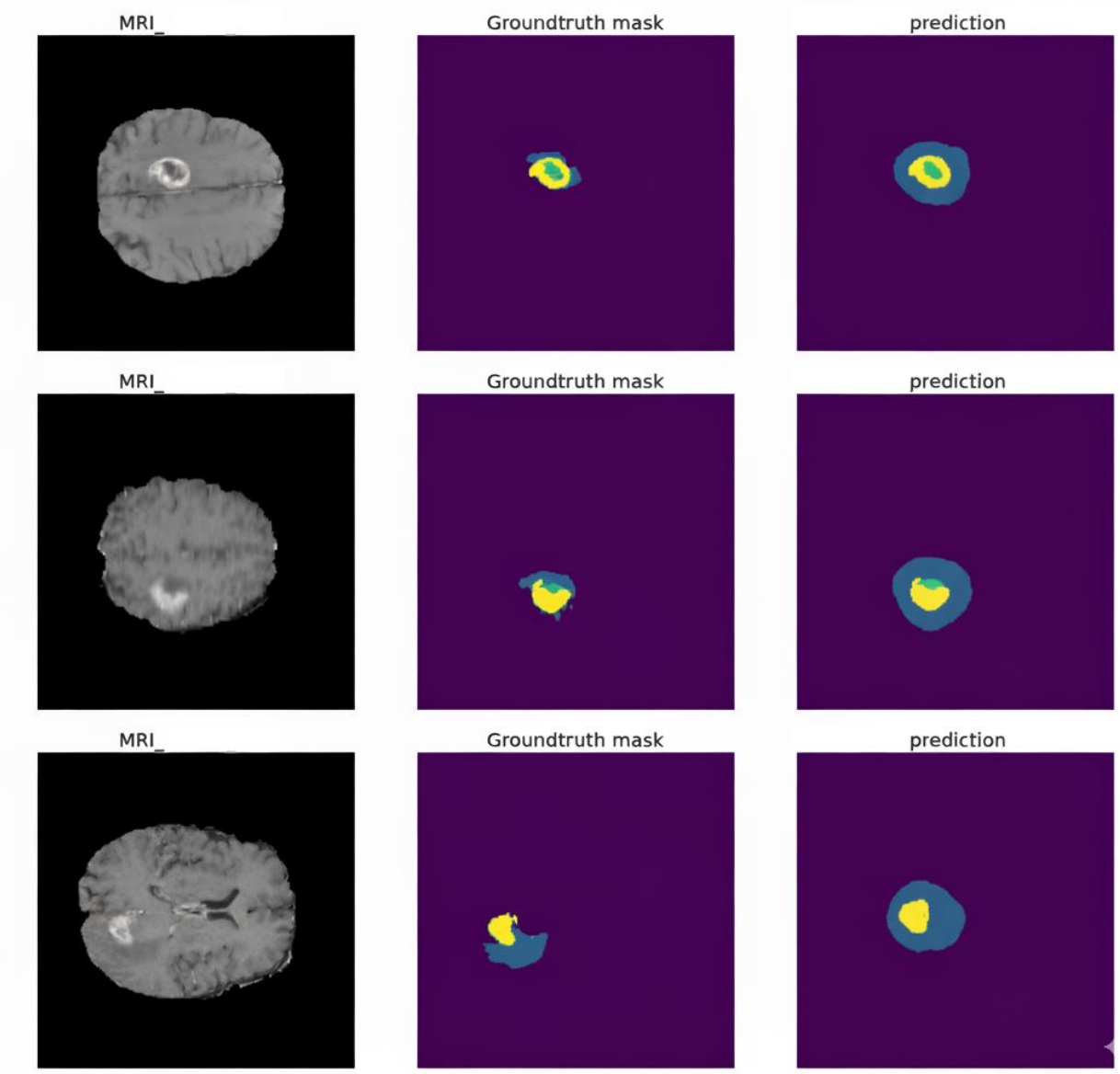


Figure 25: results obtained from Swin-UNETR + CAUSE-MLP for BraTs 2016-2017 Dataset.

5. Discussion and Conclusion

5.1 Results Discussion

The experimental results highlight clear trends regarding the influence of backbone choice, segmentation head, and dataset characteristics.

Overall, Swin-UNETR consistently outperformed DINOv2 across both the LGG and BraTS datasets, not only in terms of quantitative metrics (mIoU, mAP, Dice) but also in qualitative visualizations of tumor segmentation. Predictions obtained with DINOv2 often failed to capture tumor boundaries correctly, producing masks that were either fragmented or overly diffuse. In contrast, Swin-UNETR produced segmentations that were more coherent and anatomically plausible, reflecting its advantage as a model pretrained directly on volumetric medical data. These findings suggest that while DINOv2 demonstrates strong generalization capabilities on natural images, its transfer to MRI remains limited, likely due to the significant domain gap and the suboptimal construction of the clusterbook in Step 1, even after parameter tuning.

Regarding the segmentation heads, the differences between CAUSE-MLP and CAUSE-TR were less pronounced compared to the impact of the backbone. The transformer-based head (CAUSE-TR) showed slightly lower mIoU but achieved competitive or higher mAP values in some settings, suggesting that while it may not consistently improve spatial precision, it can better capture global consistency. However, the lightweight MLP head proved more stable overall and easier to optimize, especially on the smaller LGG dataset.

At the dataset level, the experiments do not confirmed the added difficulty of moving from the simpler, binary LGG dataset to the more complex, multi-class BraTS dataset. Performance remained similar in absolute terms, as expected, but the relative benefit of Swin-UNETR over DINOv2 remained consistent, underlining the importance of domain-specific pretraining when handling high variability and subtle tumor subregions in volumetric MRI.

Finally, it is important to acknowledge that an ablation study—systematically isolating the contribution of the mediator construction, the concept bank, and the segmentation heads—was not conducted due to time limitations. Such a study would represent a crucial next step to fully

disentangle the role of each component in driving segmentation performance and to better understand the observed differences across backbones.

Another important step here is to benchmark the results against state-of-the-art baselines.

Benchmarking in this field presents unique challenges. Unlike natural image domains, where large-scale, standardized datasets such as ImageNet or COCO have enabled fair and direct comparison across models, the biomedical domain suffers from limited data availability, strong domain-specific variability, and heterogeneous annotation protocols.

Within this constrained setting, unsupervised semantic segmentation methods remain relatively underexplored, particularly for brain tumor segmentation in datasets such as LGG or BraTS. Most state-of-the-art approaches reported in the literature either assume full supervision or adopt semi-supervised paradigms. Purely unsupervised methods such as autoencoders (AE) [68], variational autoencoders (VAE) [68], or generative approaches like f-AnoGAN [69] have been applied to BraTS, but their performance is typically limited, with Dice coefficients in the range of 31–43%. More sophisticated frameworks, such as DCTAE or CUAS [70], have pushed these results higher, achieving Dice scores of approximately 38–55%. The strongest reported method in this category, PL-BTS [71], reaches up to 79.6% Dice on BraTS2021, although its formulation is not fully aligned with the strict unsupervised clustering paradigm followed by CAUSE. These numbers highlight both the potential and the limitations of existing unsupervised pipelines: while they demonstrate progress, they also underline the methodological diversity and lack of consensus in evaluation setups, which makes direct benchmarking against CAUSE somewhat difficult.

Expanding the comparison to semi-supervised learning offers a clearer picture of the potential and limitations of our approach. Recent semi-supervised methods, such as SemiSAM+ [72], demonstrate how minimal annotation can dramatically improve segmentation quality. On BraTS 2019, this method achieves Dice scores of 68.5% with just one labeled case, increasing to 76% with five labeled cases, and ultimately reaching a fully supervised upper bound of 87.3% with all 250 labeled cases. These results showcase how even a small degree of supervision provides powerful guidance to the model, significantly closing the gap to supervised performance. Placed within this spectrum, our results, Dice scores between 0.55 (DINOv2) and 0.67 (Swin-UNETR) on BraTS, and around 0.57–0.63 on the LGG dataset are competitive for a fully unsupervised method

that requires no annotation at all. While they do not match the performance of semi-supervised pipelines, they demonstrate meaningful progress in a space where annotated data may be unavailable.

5.2 Conclusion

This thesis set out to investigate whether the semantic richness of self-supervised Vision Transformers could be effectively combined with the bias-reduction capabilities of causal inference frameworks to improve weakly supervised semantic segmentation in biomedical imaging. By adapting the CAUSE framework to MRI brain tumor segmentation and evaluating it with two distinct backbones — DINOv2 and Swin-UNETR — the work contributes both a conceptual exploration and an empirical assessment of this novel integration.

The contributions of this thesis can be summarized as follows. First, it presents a detailed analysis of the CAUSE framework, extending its application from natural images to the biomedical domain, with a focus on MRI brain tumor segmentation. Second, it evaluates the framework with two different backbones: DINOv2, a large-scale self-supervised Vision Transformer trained on 142 million natural images, and Swin-UNETR, a domain-specific model pretrained on medical imaging data. Finally, the thesis provides a comparative study of results obtained across two public MRI datasets — the Kaggle LGG dataset and BraTS 2016–2017 — under varying architectural and parameter settings.

The experimental results highlight several key insights. On the one hand, DINOv2, despite being trained exclusively on natural images, was able to provide a baseline segmentation performance with mIoU values in the low 40% range and Dice scores around 0.55–0.57. This underscores its impressive generalization capabilities and validates the hypothesis that large-scale self-supervised backbones can transfer knowledge even to domains not explicitly represented in their pretraining. However, visual inspections revealed that the tumor regions predicted with DINOv2 often lacked anatomical plausibility, suggesting that the clustering process struggles to align with subtle biomedical features when the backbone has not been exposed to medical imaging during training.

On the other hand, Swin-UNETR produced more coherent and clinically plausible segmentations. Both quantitative results (mIoU up to 46% and Dice scores around 0.63–0.67) and qualitative visualizations demonstrated a clear advantage of using a medical-pretrained backbone. This

highlights the importance of domain alignment: while general-purpose models like DINOv2 provide strong representation power, backbones pretrained on medical data better capture the structural and textural nuances of MRI, leading to more stable concept clustering and improved downstream segmentation.

These findings should be interpreted in the broader research landscape. In biomedical imaging, benchmarking remains inherently challenging: datasets are scarce, often heterogeneous across institutions, and annotations are expensive and prone to inter-observer variability. Public datasets such as LGG and BraTS represent crucial benchmarks but are still limited in scale compared to the datasets used in natural image research. Furthermore, very few unsupervised methods exist for direct comparison, and those that do are typically evaluated under different conditions or datasets. When positioned against semi-supervised methods, such as SemiSAM+ or recent multimodal SSL approaches, the performance of CAUSE-based methods lags behind supervised or weakly supervised state of the art. Yet, the annotation-free nature of CAUSE and its reliance on causal principles make it a compelling avenue for research, especially in settings where labeled data are extremely limited.

This thesis is not without limitations. Computational constraints limited the exploration of broader ablation studies, which would be necessary to disentangle the contribution of each component — from the mediator construction to the concept bank and the choice of segmentation head. Additionally, the evaluation was confined to two datasets; broader testing across modalities such as cardiac MRI, abdominal CT, or histopathology would be required to confirm generalization capabilities. Another limitation is the lack of region-specific Dice metrics (e.g., WT, TC, ET for BraTS), which would provide finer-grained insights into model behavior.

Despite these constraints, the work contributes a meaningful proof of concept. It demonstrates that causal reasoning can be embedded into self-supervised frameworks for biomedical imaging, and that the choice of backbone — general-purpose versus domain-specific — has a decisive impact on the outcome. More importantly, it highlights the potential of causality-inspired architectures to address long-standing challenges in biomedical segmentation: the scarcity of labeled data, the variability of imaging protocols, and the ambiguity of anatomical boundaries.

Looking forward, several promising directions emerge. First, ablation studies are a necessary step to isolate the contributions of mediator construction, positive/negative selection, and concept banks. Second, extending the evaluation to more recent datasets (e.g., BraTS 2018–2021, ISLES) and to other modalities would provide stronger evidence of generalization. Third, future research should explore hybridization with foundational and multimodal models, leveraging language or textual supervision to enrich the causal clustering process. Finally, emerging self-supervised models such as DINOv3 or MedMAE offer opportunities to further improve representation learning and may synergize strongly with causal frameworks like CAUSE.

In conclusion, while this work does not yet achieve state-of-the-art segmentation performance compared to fully supervised or semi-supervised models, it lays the groundwork for a promising research direction. By uniting self-supervised Vision Transformers and causal inference, it contributes a new perspective to the quest for data-efficient, annotation-free, and semantically consistent biomedical image segmentation.

References

[1] Minaee, Shervin, et al. "Image segmentation using deep learning: A survey." *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021): 3523-3542.

[2] Papanastasiou, Giorgos, et al. "Is attention all you need in medical image analysis? A review." *IEEE Journal of Biomedical and Health Informatics* 28.3 (2023): 1398-1411.

[3] Christ, Patrick Ferdinand, et al. "Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks." *arXiv preprint arXiv:1702.05970* (2017).

[4] Heller, Nicholas, et al. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge." *Medical image analysis* 67 (2021): 101821.

[5] Liu, Xiaoyu, et al. "Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation." *BioMedical Engineering OnLine* 23.1 (2024): 52.

[6] Antonelli, Michela, et al. "The medical segmentation decathlon." *Nature communications* 13.1 (2022): 4128.

[7] Turečková, Alžběta, et al. "Improving CT image tumor segmentation through deep supervision and attentional gates." *Frontiers in Robotics and AI* 7 (2020): 106.

[8] Yagis, Ekin, et al. "Deep learning for 3D vascular segmentation in hierarchical phase contrast tomography: a case study on kidney." *Scientific Reports* 14.1 (2024): 27258.

[9] Yagis, Ekin, et al. "Deep learning for 3D vascular segmentation in hierarchical phase contrast tomography: a case study on kidney." *Scientific Reports* 14.1 (2024): 27258.

[10] Thomas, Simon M., et al. "non-melanoma skin cancer segmentation for histopathology dataset." *Data in brief* 39 (2021): 107587.

[11] Jha, Debesh, et al. "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning." *Ieee Access* 9 (2021): 40496-40510.

[12] Wang, Risheng, et al. "Medical image segmentation using deep learning: A survey." *IET image processing* 16.5 (2022): 1243-1267.

[13] Zhang, Chen, Xiangyao Deng, and Sai Ho Ling. "Next-Gen medical imaging: U-Net evolution and the rise of transformers." *Sensors* 24.14 (2024): 4668.

[14] Liu, Xiangbin, et al. "A review of deep-learning-based medical image segmentation methods." *Sustainability* 13.3 (2021): 1224.

[15] Hesamian, Mohammad Hesam, et al. "Deep learning techniques for medical image segmentation: achievements and challenges." *Journal of digital imaging* 32.4 (2019): 582-596.

[16] Bougourzi, Fares, and Abdenour Hadid. "Recent advances in medical imaging segmentation: A survey." *arXiv preprint arXiv:2505.09274* (2025).

[17] Weese, Jürgen, and Cristian Lorenz. "Four challenges in medical image analysis from an industrial perspective." *Medical image analysis* 33 (2016): 44-49.

[18] Upadhyay, Ashwini Kumar, and Ashish Kumar Bhandari. "Advances in deep learning models for resolving medical image segmentation data scarcity problem: a topical review." *Archives of Computational Methods in Engineering* 31.3 (2024): 1701-1719.

[19] Guan, Hao, and Mingxia Liu. "Domain adaptation for medical image analysis: a survey." *IEEE Transactions on Biomedical Engineering* 69.3 (2021): 1173-1185.

[20] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

[21] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Cham: Springer international publishing, 2015.

[22] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." *International workshop on deep learning in medical image analysis*. Cham: Springer International Publishing, 2018.

[23] Oktay, Ozan, et al. "Attention u-net: Learning where to look for the pancreas." *arXiv preprint arXiv:1804.03999* (2018).

[24] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." *International conference on medical image computing and computer-assisted intervention*. Cham: Springer International Publishing, 2016.

[25] Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature methods* 18.2 (2021): 203-211.

[26] Azad, Reza, et al. "Medical image segmentation review: The success of u-net." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[27] Du, Getao, et al. "Medical image segmentation based on U-net: A review." *Journal of Imaging Science & Technology* 64.2 (2020).

[28] Chen, Zhaozheng, and Qianru Sun. "Weakly-supervised semantic segmentation with image-level labels: from traditional models to foundation models." *ACM Computing Surveys* 57.5 (2025): 1-29.

[29] Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[30] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

[31] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018.

[32] Omeiza, Daniel, et al. "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models." *arXiv preprint arXiv:1908.01224* (2019).

[33] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[34] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[35] Wang, Wenxuan, et al. "Transbts: Multimodal brain tumor segmentation using transformer." *International conference on medical image computing and computer-assisted intervention*. Cham: Springer International Publishing, 2021.

[36] Xie, Yutong, et al. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation." *International conference on medical image computing and computer-assisted intervention*. Cham: Springer International Publishing, 2021.

[37] Chen, Jieneng, et al. "Transunet: Transformers make strong encoders for medical image segmentation." *arXiv preprint arXiv:2102.04306* (2021).

[38] Heidari, Moein, et al. "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023.

[39] Chi, Jianning, et al. "Hybrid transformer UNet for thyroid segmentation from ultrasound scans." *Computers in Biology and Medicine* 153 (2023): 106453.

[40] Sun, Guanqun, et al. "DA-TransUNet: integrating spatial and channel dual attention with transformer U-net for medical image segmentation." *Frontiers in Bioengineering and Biotechnology* 12 (2024): 1398237.

[41] Balestriero, Randall, et al. "A cookbook of self-supervised learning." *arXiv preprint arXiv:2304.12210* (2023).

[42] Kotei, Evans, and Ramkumar Thirunavukarasu. "A systematic review of transformer-based pre-trained language models through self-supervised learning." *Information* 14.3 (2023): 187.

[43] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[44] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).

[45] Zhou, Jinghao, et al. "ibot: Image bert pre-training with online tokenizer." *arXiv preprint arXiv:2111.07832* (2021).

[46] Bao, Hangbo, et al. "Beit: Bert pre-training of image transformers." *arXiv preprint arXiv:2106.08254* (2021).

[47] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[48] Assran, Mahmoud, et al. "Self-supervised learning from images with a joint-embedding predictive architecture." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[49] Siméoni, Oriane, et al. "DINOv3." *arXiv preprint arXiv:2508.10104* (2025).

[50] Weng, Xi, et al. "Clustering Properties of Self-Supervised Learning." *arXiv preprint arXiv:2501.18452* (2025).

[51] Su, Qing, et al. "Flsl: Feature-level self-supervised learning." *Advances in Neural Information Processing Systems* 36 (2023): 6568-6581.

[52] Valentinos Pariza, Mohammadreza Salehi, Gertjan J Burghouts, Francesco Locatello, and Yuki M Asano. Near, far: Patch-ordering enhances vision foundation models' scene understanding. In ICLR, 2025.

[53] Ziegler, Adrian, and Yuki M. Asano. "Self-supervised learning of object parts for semantic segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022

[54] Hamilton, Mark, et al. "Unsupervised semantic segmentation by distilling feature correspondences." *arXiv preprint arXiv:2203.08414* (2022).

[55] Tucci, Robert R. "Introduction to Judea Pearl's Do-Calculus." *arXiv preprint arXiv:1305.5506* (2013).

[56] Zhang, Kexuan, et al. "Causal reasoning in typical computer vision tasks." *Science China Technological Sciences* 67.1 (2024): 105-120.

[57] Liu, Yang, et al. "Causal reasoning meets visual representation learning: A prospective study." *Machine Intelligence Research* 19.6 (2022): 485-511.

[58] Niu, Yulei, et al. "Counterfactual vqa: A cause-effect look at language bias." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

[59] He, Xuehai, et al. "Cpl: Counterfactual prompt learning for vision and language models." *arXiv preprint arXiv:2210.10362* (2022).

[60] Junho Kim et al. "Causal Unsupervised Semantic Segmentation". arXiv:2310.07379 (2023).

[61] Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., & Brundage, M. (2021). Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

[62] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).

[63] Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., & Zhang, S. (2022, September). Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 528-538). Cham: Springer Nature Switzerland.

[64] Xing, Z., Wan, L., Fu, H., Yang, G., & Zhu, L. Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv 2023. *arXiv preprint arXiv:2303.10326*.

[65] Seong, Hyun Seok, et al. "Leveraging hidden positives for unsupervised semantic segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

[66] Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., & Peng, W. (2023). Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. *arXiv preprint arXiv:2312.02366*.

[67] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop* (pp. 272-284). Cham: Springer International Publishing.

[68] Baur, C., Denner, S., Wiestler, B., Navab, N., & Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Medical image analysis*, *69*, 101952.

[69] Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., & Schmidt-Erfurth, U. (2019). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, *54*, 30-44.

[70] Silva-Rodríguez, J., Naranjo, V., & Dolz, J. (2022). Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, *80*, 102526.

[71] Zhang, X., Ou, N., Liu, C., Zhuo, Z., Liu, Y., & Ye, C. (2023). Unsupervised Brain Tumor Segmentation with Image-based Prompts. *arXiv preprint arXiv:2304.01472*.

[72] Zhang, Y., Lv, B., Xue, L., Zhang, W., Liu, Y., Fu, Y., ... & Qi, Y. (2025). SemiSAM+: Rethinking Semi-Supervised Medical Image Segmentation in the Era of Foundation Models. *arXiv preprint arXiv:2502.2074*.