Cerebrovascular Pathology Segmentation Using Weakly Supervised Deep Learning Methods

Amirhossein Rasoulian

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

December 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Amirhossein Rasoulian**

Entitled: **Cerebrovascular Pathology Segmentation Using Weakly Supervised Deep Learning Methods**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Charalambos Poullis

_____ External Examiner
Dr. Eugene Belilovsky

_____ Examiner
Dr. Charalambos Poullis

_____ Supervisor
Dr. Yiming Xiao

Approved by _____
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2023

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Cerebrovascular Pathology Segmentation Using Weakly Supervised Deep Learning Methods

Amirhossein Rasoulian

Intracranial hemorrhage (ICH) and unruptured intracranial aneurysm (UIA) are two important cerebrovascular diseases that require prompt and precise diagnosis for effective treatment and improved survival rates. While deep learning (DL) techniques have emerged as the leading approach for medical image analysis and processing, the most commonly employed supervised learning often requires large, high-quality annotated datasets that can be costly to obtain, particularly for pixel/voxel-wise image segmentation. To address this challenge and meet the need in cerebrovascular care, we proposed and validated three novel weakly supervised segmentation methods for ICH using categorical labels and for UIA with coarse image segmentation. For ICH, we first introduced a framework to segment the lesion based on a hierarchical combination of self-attention maps obtained from a Swin transformer, which was trained only for ICH detection, achieving a Dice score of 0.407. Subsequently, by employing novel head-wise gradient-weighing of self-attention maps in the same setup, we further improved the mean Dice score to 0.444 for ICH segmentation. Our method that only relies on categorical labels showed comparable performance against popular fully supervised methods, such as UNet and Swin-UNETR. Finally, for UIA segmentation, we achieved a Dice score of 0.68 and a 95% Hausdorff distance of $\sim$0.95 mm by proposing a new 3D focal modulation UNet, called FocalSegNet. This novel DL architecture was trained with coarse manual segmentation, providing an initial segmentation of aneurysms, which was then refined using dense conditional random field (CRF) post-processing. Our proposed methods explored new avenues using weak labels to mitigate a key bottleneck in medical DL with excellent performance and showcased their promising potential in addressing challenging medical image segmentation tasks.

Acknowledgments

First and foremost, I would like to express my deep gratitude to my supervisor, Dr. Yiming Xiao for giving me the invaluable opportunity to work in HealthX Lab. He has always pushed me, given me new ideas, and helped me with any kind of issues I countered even outside the academia. I cannot recall any moment that I messaged him, and the response took longer than one minute. His guidance has been instrumental in shaping both my academic and professional journey.

I extend my gratitude to Dr. Aiman Hanna for consistently assisting me in securing teaching assistant positions each semester. This support not only alleviated financial pressures but also enriched my experiences significantly.

Throughout this journey, the combination of living far from home, studying, working, and engaging in research has endowed me with priceless experiences.

Lastly, heartfelt thanks to my family and friends for their unwavering encouragement, advice, and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What is a cerebrovascular disease?

Cerebrovascular disease refers to a group of medical conditions that affect the blood vessels in the brain, such as strokes, aneurysms, and vascular malformations. A stroke is a sudden and often severe neurological event that occurs when there is an interruption in the brain's blood supply. This interruption can be due to:

- **Ischemic Stroke:** blockage of brain blood vessels, typically by a blood clot or plaque, causing reduced blood flow to a part of the brain and the death of brain cells.
- **Hemorrhagic Stroke:** rupture of a blood vessel within the brain, leading to bleeding inside the brain (intracerebral hemorrhage) or in the space surrounding the brain (subarachnoid hemorrhage).

Strokes are a significant cause of morbidity and mortality worldwide and can have serious consequences for brain function. According to World Stroke Organization (WSO) statistics, more than 12.2 million new strokes occur worldwide every year, and there is the annual demise of six and a half million individuals. In the global population aged 25 and above, one out of every four individuals is expected to experience a stroke during their lifetime. Intracranial Aneurysm constitutes another category of cerebrovascular disorder, involving an abnormal localized bulging or ballooning of the blood vessels in the brain due to the weakened integrity of the blood vessel walls. It has

the potential to result in a critical condition known as subarachnoid hemorrhage in case it ruptures. While frequently without noticeable symptoms, larger aneurysms can give rise to manifestations such as severe headaches, visual impairments, neurological deficiencies, and various other complications [1]. In this thesis, we opted to focus on hemorrhagic strokes, which are less prevalent than ischemic strokes, but they are deadlier. Additionally, the document explores the topic of Intracranial Aneurysms, which have the potential to result in subarachnoid hemorrhage.

1.2 Challenges

Hemorrhagic strokes are diagnosed using CT or MRI imaging in clinical settings. Timely and accurate detection is crucial because rapid intervention is vital for minimizing brain damage and improving patient outcomes. However, manually analyzing these images for signs of hemorrhagic stroke is time-consuming and requires expertise. Automatic detection methods are desired to expedite the process and enable faster, more accurate diagnosis, which is particularly critical in stroke cases where every minute counts. On the other hand, most deep learning-based medical image segmentation methods rely heavily on training with large and meticulously annotated datasets. However, this presents a formidable challenge in the field, as gathering such datasets is a time-consuming and expensive process. It necessitates the expertise of medical professionals to accurately label images, and accumulating a substantial number of diverse cases takes considerable effort and resources.

1.3 Proposed solution

In response to the challenge of insufficient high-quality annotated images for training deep learning models, our approach involves the utilization of weakly supervised learning techniques. In this methodology, the model is trained using less precise labels, but it is equipped with the capability to generate accurate and high-quality predictions. To elaborate further, our method utilizes image categorical labels or bounding boxes during training, and at the testing stage, it leverages these techniques to produce voxel-wise segmentation, ensuring the production of precise and detailed results.

1.4 Thesis outline

This thesis is organized in 6 chapters. Chapter 2 offers the foundational knowledge necessary for comprehending the remainder of the thesis. It covers the topics of medical image segmentation, weakly supervised learning strategy, and transformer models. In Chapter 3, we propose a technique that does intracranial hemorrhage (ICH) segmentation, employing a hierarchical combination of attention maps derived from a Swin transformer trained on categorical labels. Moving on, in Chapter 4, we enhance the previously introduced model by integrating the concept of class-activation mapping and combining it with attention maps, featuring a specialized head-wise gradient-infused self-attention mapping technique for precise ICH lesion segmentation. Chapter 5 shifts the focus to the weakly supervised segmentation of intracranial aneurysms. Here, we propose a novel segmentation model using focal modulation, trained on weak labels represented as spheres encompassing the target. This model provides an initial coarse segmentation, subsequently refined through conditional random field post-processing to achieve voxel-wise aneurysm segmentation. Finally, in Chapter 6, we conclude the thesis and discuss potential future improvements.

Chapter 2

Background and Literature Review

In this chapter, we start by elucidating the concept of medical image segmentation, explaining diverse medical image modalities, prevalent pre-processing techniques, and metrics to evaluate the performance of models. Following this, we introduce the deep learning methodologies serving as the foundational architecture throughout this thesis. We then navigate into the domain of weakly supervised learning, and we provide a detailed explanation of specific weakly supervised learning methods applied in this thesis for the segmentation of hemorrhagic strokes. Lastly, we undertake a thorough review of related papers employing weakly supervised techniques for the medical image segmentation task.

2.1 Medical Image segmentation

Image segmentation involves dividing an image into disjoint homogeneous regions based on features like intensity, spatial texture, and geometric shapes. This process extracts valuable information by ensuring similarity within each area while highlighting clear differences between them. Medical images play a pivotal role in clinical diagnosis, and automatic segmentation methods offer consistent accuracy and enhanced speed compared to manual approaches. Applying computer-assisted methods to segment medical images assists doctors in evaluating tumor/lesion size, quantifying treatment effects, and significantly reducing their workload.

2.1.1 Introduction to Medical Imaging

Medical imaging is the procedure of capturing images of the internal structures of the body for clinical examination, medical interventions, and visual illustrations of the functioning of certain organs or tissues. Its primary objective is to unveil hidden internal structures beneath the skin, facilitating the diagnosis and treatment of various diseases. Advancements in medical treatment have resulted in the extensive adoption of imaging technologies, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), X-ray, and ultrasound in clinics. Each modality possesses distinctive strengths and weaknesses, designed for specific applications. In contrast to RGB natural images, medical images are typically grayscale images with diverse intensities. They come with specific metadata for each image, including patient information and an affine matrix used to derive real-world coordinates of image voxels. Consequently, these images are stored in formats like NIfTI, DICOM, and MINC, which are specifically designed to store medical data. The predominant imaging modalities employed for brain imaging include CT and MRI, both of which will be discussed in this section. To provide a comprehensive overview, a detailed comparison between these modalities is outlined in Table 2.1, complemented by illustrative examples of their images depicted in Fig. 2.1.

Magnetic Resonance Imaging

An MRI machine utilizes the interaction of magnetic fields and protons in biological tissues to produce 3D/4D images. The scanner polarizes and excites hydrogen nuclei within water molecules in body tissue by emitting a radio frequency (RF) pulse at the resonant frequency of hydrogen atoms. This RF pulse, delivered through radio-frequency antennas, causes protons to absorb the energy, altering their alignment with the primary magnetic field. Once the RF pulse is deactivated, the protons return to their original alignment and emit radio waves. These emitted signals are then detected and spatially encoded to reconstruct detailed images of the body. Fig. 2.1 (A) illustrates an example of brain MRI image.

Computed Tomography

In CT imaging, a rotating beam of X-rays surrounds the object under examination, and sensitive radiation detectors capture the X-rays after penetrating the object from multiple angles. The collected data is then analyzed by a computer, which utilizes mathematical principles, particularly the Radon transform, to construct a detailed 3D image of the object and its internal structures. To mitigate potential health risks, it is crucial to restrict the frequency of repeated CT scans, as they subject the patient to ionizing radiation emitted during X-ray projections. Fig. 2.1 (B) shows an example of brain CT image.

Angiography Imaging

Angiography images offer a detailed visualization of blood vessels throughout the body, facilitating a comprehensive assessment of their structure and function. Widely employed for evaluating blood flow and identifying conditions like blockages and aneurysms, these images play a crucial role in planning medical interventions such as surgeries or angioplasty. The imaging involves the injection of a contrast agent into the blood vessels, and subsequent utilization of MRI or CT techniques to capture detailed images, revealing any abnormalities present in the vasculature. Fig. 2.1 (C) and Fig. 2.1 (D) depicts examples of brain MR Angiography (MRA) and CT Angiography (CTA) images, respectively.

Table 2.1: Comparison of MRI and CT

MRI	**CT**
Non-invasive and does not involve ionizing radiation, making it safer for repeated use.	Exposes patients to ionizing radiation, raising the risk of developing cancer.
Offers excellent soft tissue contrast for imaging organs, muscles, and the nervous system.	Provides good contrast for bones but is less effective in distinguishing between different soft tissues.
Scans can take longer, potentially limiting use for patients who have difficulty remaining still.	Relatively quick imaging, valuable in emergency situations.
Certain metallic implants can interfere with imaging.	Versatile and applicable to various body parts, including the head, chest, abdomen, and musculoskeletal system.
Equipment and procedures tend to be more expensive, limiting availability in some settings.	Widely available in hospitals and medical facilities, ensuring accessibility.

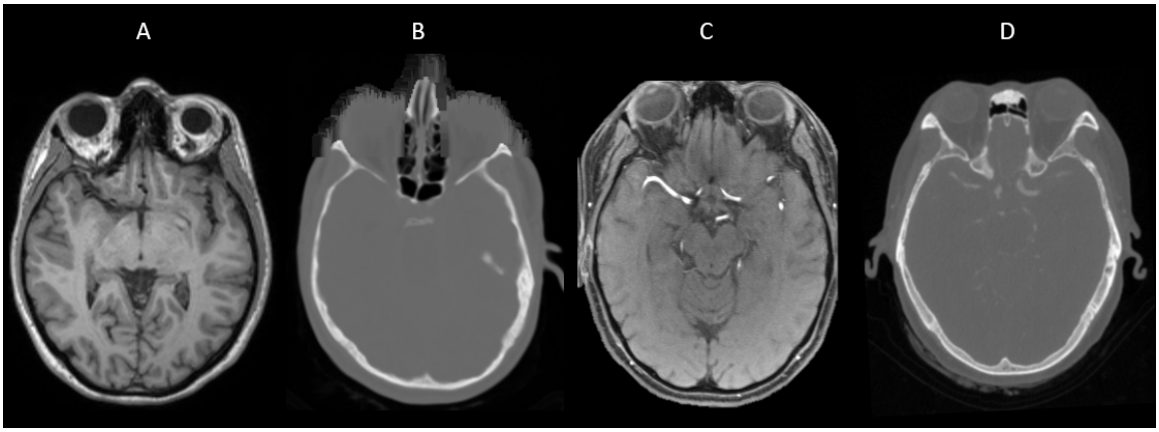

Figure 2.1: Visualization example of different medical image modalities: A) MRI. B) CT. C) MRA. D) CTA.

2.1.2 Medical Image Pre-processing

Image Registration

Image registration, involving image fusion, matching, and warping, is the process of spatially aligning multiple images. In the context of brain images, image fusion involves merging brain images of a population to create a standard average image, commonly known as an atlas; image matching involves determining a spatial transformation that aligns one brain image with a brain atlas (or another brain image), enabling the extraction of landmark coordinates from the atlas on the new brain image; and image warping involves applying a geometric transformation, such as registering a new brain image to a brain atlas, ensuring spatial standardization for comprehensive analysis [2]. While this thesis does not dive deep into the details of medical image registration techniques, it is still relevant in pre-processing of the data used in the experiments. Classic image registration typically begins with the initialization of a transformation matrix, followed by the iterative reduction of the energy/loss function to achieve the optimal alignment.

Resampling

In medical imaging, the accurate size of each voxel (three-dimensional pixel) is critical as it directly influences the assessment of lesions, pathologies, and disease progression. While the image dimension indicates the number of voxels in each dimension of the image grid, the physical resolution

specifies the actual size of each voxel in millimeters. For instance, in a $1 \times 1 \times 1mm$ resolution image, if a lesion encompasses 10 voxels, its true volume is $10mm^3$. The resolution of medical images varies based on factors like imaging modality, protocol, and scanner type. Consequently, a common pre-processing step in medical images involves resampling all images to a standardized template space with a unified resolution, ensuring comparability. Resampling entails interpolating pixel values to create a new pixel grid with adjusted locations and/or resolutions.

MRI Bias Field Correction

MRI bias field, also known as intensity inhomogeneity, refers to low frequency and nonanatomic intensity variations within the image domain of the same tissue. This artifact can arise from factors like imaging instrumentation (e.g., radio-frequency nonuniformity, static field inhomogeneity). It is characterized by significant intensity variations among pixels within the same tissue and notable overlap in intensity values between pixels of different tissues, which can affect the accuracy of MR image analysis. Fig. 2.2 (A) visually illustrates this artifact with two real MR images displaying the impact of the MRI bias field.

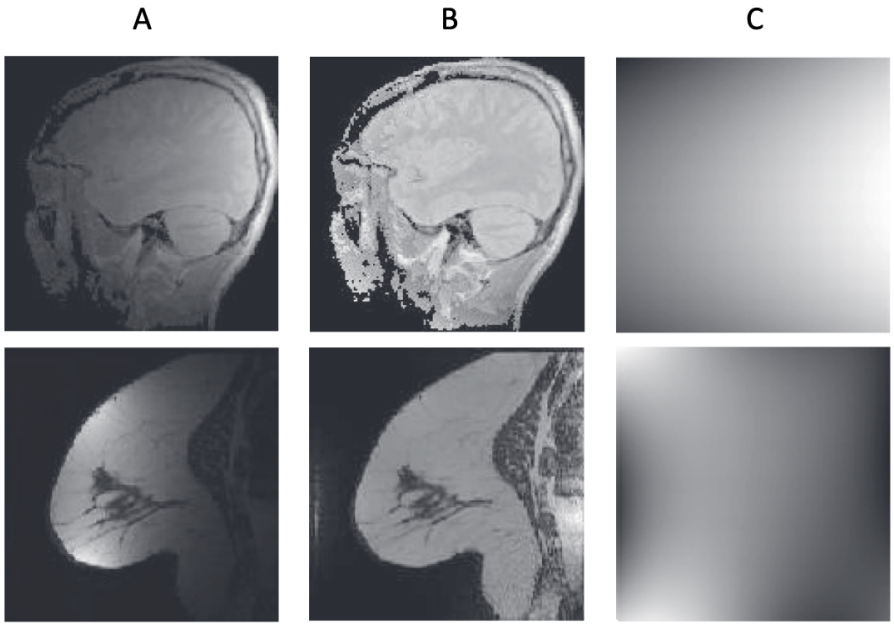

Figure 2.2: Examples of MRI bias field artifact: A) original images. B) corrected images. C) estimated inhomogeneity maps. [3]

To address the inhomogeneity artifact, the image formation model is assumed as follows:

$$v(x) = u(x)f(x) + n(x) \tag{1}$$

where v and u denote the original corrupted and corrected images, respectively; f represents the bias field, and n signifies the noise. Typically, the noise in the equation is neglected, leading to a simplified model:

$$\log v(x) = \log u(x) + \log f(x) \tag{2}$$

Subsequently, the bias field is estimated, and the corrected image is obtained through subtraction in the log-domain. The state-of-the-art method for bias field correction is N4ITK [4]. It is an iterative optimization process that aims to estimate a B-spline approximator of a smooth multiplicative field that maximizes the high-frequency content within the distribution of tissue intensity.

CT Windowing

CT image intensities are expressed in Hounsfield Units (HU), named after the CT scan inventor, reflecting the density and characteristics of different tissues. The HU scale ranges from -1000 HU for air to over 3000 for metals. Due to this broad dynamic intensity range in CT scans, a windowing technique is applied to enhance contrast for specific structures, selecting a window of HU values corresponding to that structure. Eq. 3 illustrates how the windowing function W generates new intensities for Image x.

$$\min = \text{WL} - \frac{\text{WW}}{2} \tag{3a}$$

$$\max = \text{WL} + \frac{\text{WW}}{2} \tag{3b}$$

$$W(x) = \begin{cases} \min & \text{if } x < \min \\ x & \text{if } \min \leq x \leq \max \\ \max & \text{if } \max < x \end{cases} \tag{3c}$$

The Window Level (WL) denotes the window center or midpoint HU value represented in the window setting, while the Window Width (WW) measures the range of HU values to be preserved in a CT image. Adjusting these windowing parameters allows for optimal contrast for a specific tissue, as illustrated in Fig. 2.3.

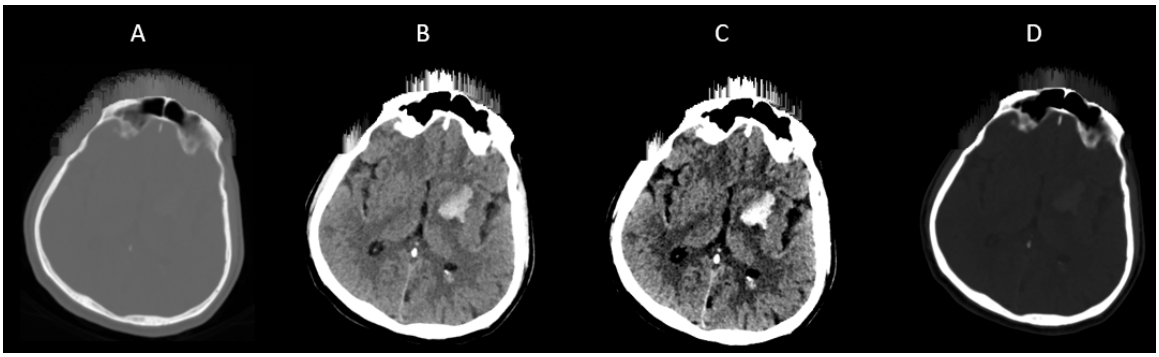

Figure 2.3: Examples of different brain CT windows: A) original CT image. B) brain window (WL=40, WW=80). C) stroke window (WL=40, WW=40). D) bone window (WL=400, WW=1000).

Intensity Normalization

The range of pixel intensities in CT and MRI images can vary based on several factors, including the imaging protocol and the specific characteristics of the scanner used. Thus, it is essential to employ intensity normalization techniques during pre-processing to ensure that all images are on a comparable scale and no single image dominates others in the segmentation model training or statistical analyses. As depicted in Eq. 4, two common and simple methods, min-max normalization $N(x)$ and Z normalization $Z(x)$, can be employed on the input image x to standardize its intensity range for deep learning purposes.

$$N(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4a}$$

$$Z(x) = \frac{x - \text{mean}(x)}{\text{std}(x)} \tag{4b}$$

$N(x)$ scales intensities to the range of $[0, 1]$, while $Z(x)$ centers the new intensities around zero with a standard deviation of one. In addition to these methods, histogram matching-based techniques can also be employed to further adjust the intensity profile of the target image to help reduce the contrast

differences across different scanners.

2.1.3 Model Overfitting and Data Augmentation

Overfitting in a model occurs when it memorizes the training set instead of understanding the underlying function that maps input images to target classes. Building segmentation models that generalize well requires a substantial amount of ground truth data with good diversity to prevent overfitting. However, obtaining high-quality ground truth data in the medical image analysis field is time-consuming, expensive, and heavily reliant on human input. Privacy regulations further prevent the availability of publicly accessible datasets. Additionally, the majority of manually annotated image sets are imbalanced, where certain classes have only a few examples, causing the model to disregard them.

Nalepa et al. [5] conducted a comprehensive review of different data augmentation techniques employed in the literature to address the limitations posed by restricted medical training sets. These techniques involve generating synthetic training examples through various transformations like affine image modifications, elastic transformations, pixel-level changes, and other methods, artificially expanding the size of training sets. These approaches significantly improve the generalization abilities of segmentation models, as illustrated in Figure 2.4, which demonstrates these augmentations on a brain MRI.

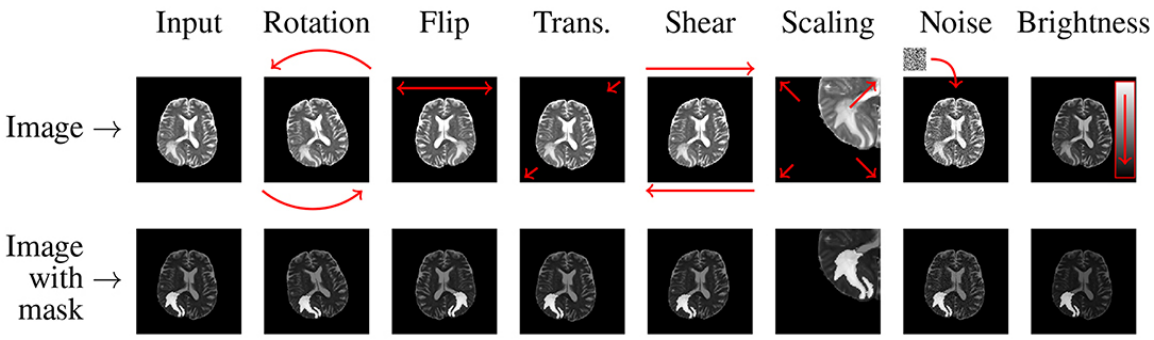

Figure 2.4: An example of medical image augmentation. [5]

2.1.4 Evaluation Methods

To assess the model's efficacy in accurately delineating pathology, tumors, or lesions, various metrics are employed to provide a quantitative comparison across different models. Throughout this thesis, we primarily utilize three segmentation metrics: Dice coefficient, Intersection over Union (IoU), and Hausdorff distance (HD). The mathematical representation of these metrics is detailed in Eq. 5.

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \tag{5a}$$

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{5b}$$

$$HD(P, G) = \max\left(\max_{p \in P} \min_{g \in G} \|p - g\|_2, \max_{g \in G} \min_{p \in P} \|g - p\|_2\right) \tag{5c}$$

The Dice coefficient and IoU, both ranging from 0 to 1 where 0 signifies no overlap and 1 indicates full intersection, serve as region-based metrics. They quantify the overlap ratio between the model's predictions (P) and the corresponding ground truth (G). In addition to those, HD is a distance-based metric that assesses the proximity of the prediction to the ground truth. The numerical value of HD is computed as the maximum Euclidean distance from any point in P to the nearest point in G, and vice versa.

2.2 Deep-Learning-Based Segmentation

Traditional image segmentation methods, such as threshold-based [6], region-based [7], and edge detection-based [8] methods, rely on digital image processing and mathematics. While they offer fast segmentation, the accuracy in capturing details is limited. In contrast, deep learning-based segmentation methods have achieved significant advancements, surpassing traditional methods in accuracy and effectiveness.

Training a Neural Network model (NN) is formulated as a sequential process, as expressed by

the equations in (6).

$$Y = \mathcal{F}(X|\theta) \tag{6a}$$

$$L = \mathcal{L}(\hat{Y}, Y) \tag{6b}$$

$$\theta_k = \theta_{k-1} - \epsilon \frac{\partial L}{\partial \theta} \tag{6c}$$

Initially, the model $\mathcal{F}$, characterized by parameters θ, undergoes random initialization. Subsequently, at each iteration, the model processes the input X to yield a prediction Y. The discrepancy between this prediction and the expected output $\hat{Y}$ is quantified by the loss function $\mathcal{L}$. The objective of the training process is to minimize this loss, thereby aligning the model's predictions more closely with the desired outputs. This optimization is achieved through iterative updates of the model's parameters using the gradient descent method, as delineated in (6c). The learning rate, denoted by ϵ, plays a crucial role in determining the step size in the direction of the gradient of the loss function with respect to the model's weights, facilitating the convergence of the model towards an optimal configuration [9].

The very basic block of a NN is a Fully Connected (FC) Layer, wherein each element of the input vector $x \in \mathbb{R}^m$ is connected to every element of the output vector $y \in \mathbb{R}^n$:

$$y_i = \mathcal{G}(\sum_{j}^{m} w_{i,j} x_j + b_i); \quad X = \{x_1, x_2, ..., x_m\}, Y = \{y_1, y_2, ..., y_n\}, \quad \forall i, j; w_{i,j} \in \theta \tag{7}$$

Here, $\mathcal{G}$ denotes an activation function that imparts non-linearity to the model $\mathcal{F}(\theta)$, and b represents a bias term. Using a stack of these FC layers, more elaborate functions can be modeled. However, its full connectivity introduces two primary drawbacks. Firstly, the abundance of parameters demands a substantial number of training samples for effective training. Secondly, it lacks invariance to translation, scale, rotation, and other transformations, limiting its ability to robustly generalize across variations in the input data.

By the remarkable success of AlexNet [10] in utilizing a deep Convolutional Neural Network

(CNN), the past decade has witnessed the dominance of CNN as the predominant choice for computer vision tasks [11]. Distinguished by its convolutional structures, CNN excels in feature extraction from data. In contrast to FC architectures, each output feature is no longer intricately connected to all elements of the input; rather, it is locally linked to a small subset, resulting in a reduction of parameters. This not only accelerates convergence, but also imparts transformation invariance to the network, enhancing its adaptability across various spatial transformations like translation and rotation.

2.2.1 U-Net

First introduced by Ronneberger et al. [12], the U-Net model was built upon the fully convolutional network to do the semantic segmentation. It is a U-shaped architecture (see Fig 2.5) consisting of a contracting path (left side) with repeated convolutions and downsampling, and an expansive path (right side) with upsampling and convolutions. The contracting path extracts features, while the expansive path recovers spatial information. The final layer uses 1×1 convolution to map features to the desired classes.

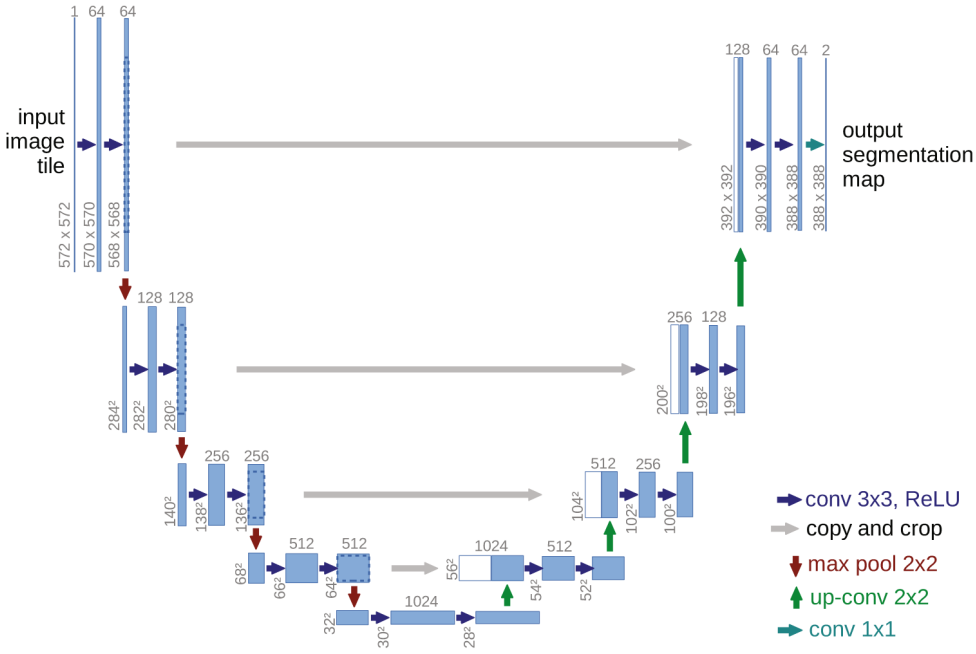

Figure 2.5: U-Net architecture. [12]

The feature extraction path, known as the encoder, holds a crucial role in segmentation models by capturing essential features from the high-dimensional input image and embedding them into a low-dimensional latent-space vector. The expansive branch (also called the decoder) utilizes these features to reconstruct the segmentation map, making the encoder performance crucial for overall network output. In the following sections, we present state-of-the-art encoders employed throughout this thesis.

2.2.2 Vision Transformer

The Transformer model, pioneered by Vaswani et al. [13], revolutionized natural language processing (NLP) by outperforming traditional sequential models like Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN). Entirely based on the self-attention mechanism, Transformer empowers each sequence element to simultaneously focus and weigh different parts of the input sequence, effectively capturing intricate relationships and long-range dependencies in the data. Inspired by the success of Transformers in NLP, Dosovitskiy et al. [14] introduced Vision Transformer (ViT) by directly applying a standard Transformer to images with minimal adjustments. This involved treating image patches similarly to tokens (words) in NLP, as illustrated in Fig 2.6.

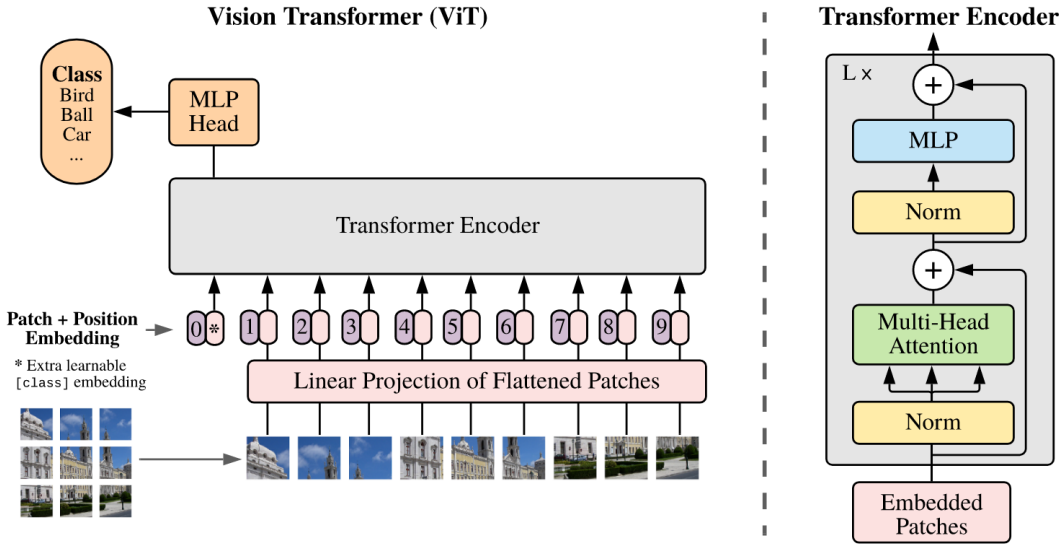

Figure 2.6: Vision Transformer architecture. [14]

The detailed operations of ViT architecture are also outlined in Eq 8:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; ...; x_p^N E] + E_{pos}, \qquad E \in \mathbb{R}^{(P^2 \cdot C) \times D}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{8a}$$

$$z_l^{'} = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \qquad l = 1...L \tag{8b}$$

$$z_l = \text{MLP}(\text{LN}(z_l^{'})) + z_l^{'}, \qquad l = 1...L \tag{8c}$$

$$y = \text{LN}(z_L^0) \tag{8d}$$

In this architecture, first, the input image $x \in \mathbb{R}^{H \times W \times C}$ is reshaped to a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) and (P, P) are the resolutions of original and patch images, respectively; C is the number of channels, and $N = HW/P^2$ is the resulting number of patches. Then, patches are linearly embedded by E, resulting in D dimensional embedding tokens. A learnable embedding $z_0^0 = x_{class}$ is also prepended to patch tokens, whose state at the output (z_0^L) determines the prediction class y (Eq. 8d). Since self-attention computes attention scores based on pairwise relationships between tokens, without positional information, a positional embedding E_{pos} is also added to the patch embeddings to retain positional information of tokens in the sequence (Eq. 8a). Next, these embeddings are passed through L Transformer layers consisting of multiheaded self-attention (MSA) and multi-layer perceptron (MLP) blocks (Eq. 8b, 8c), with layer normalization (LN) applied before each block and residual connections after each block. Ultimately, an image class is predicted using the representation z_0^L through a dedicated classification head [14].

Multihead Self-Attention

The standard **qkv** self-attention (SA, [13]) operation is shown in Eq. 9:

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv}, \qquad \mathbf{U}_{qkv} \in \mathbb{R}^{D \times 3D_h}, [\mathbf{q}, \mathbf{k}, \mathbf{v}] \in \mathbb{R}^{N \times D} \tag{9a}$$

$$A = \text{softmax}\left(\mathbf{q}\mathbf{k}^T / \sqrt{D_h}\right), \qquad A \in \mathbb{R}^{N \times N} \tag{9b}$$

$$SA(\mathbf{z}) = A\mathbf{v} \tag{9c}$$

$$MSA(\mathbf{z}) = [SA_1(\mathbf{z}); SA_2(\mathbf{z}); ...; SA_k(\mathbf{z})] \tag{9d}$$

First, the self-attention unit linearly maps the input sequence $z \in \mathbb{R}^{N \times D}$ into query q, key k, and value v vectors (Eq. 9a), where N is the number of D dimensional tokens. Then, it calculates a weighted sum over all v (Eq. 9c). The attention weights A_{ij} are determined by the pairwise dot-product similarity between two elements of the sequence using their corresponding q_i and k_j representations (Eq. 9b). As shown in Eq. 9d, Multihead self-attention (MSA) is an extension of SA where k self-attention operations, known as "heads", are executed in parallel, and their concatenated results are projected.

2.2.3 Swin Transformer

In the standard Transformer model [13], global self-attention (Eq. 9) computes the relationships between each token and all others, resulting in quadratic complexity with respect to the number of tokens (Eq. 10a). Directly applying this approach to vision, as in the Vision Transformer [14], is inefficient due to the significantly higher resolution of pixels in an image compared to words in text. To address this, Liu et al. introduced the Shifted Window (Swin) Transformer [15]. Swin Transformer divides the C-channel input image with the resolution of (h, w) into non-overlapping windows of size (M, M) patches and computes Window MSA (W-MSA) locally within each window. The fixed number of patches in each window results in linear complexity relative to the image size (Eq. 10b). Performing self-attention globally is often impractical for a large number of patches (h, w), whereas window-based self-attention is scalable.

$$\mathcal{O}(\text{MSA}) = 3hwC^2 + 2(hw)^2C, \tag{10a}$$

$$\mathcal{O}(\text{W-MSA}) = 3hwC^2 + 2(M)^2hwC, \tag{10b}$$

Shifted Window Approach

To allow for cross-window connections, Swin Transformer's design involves shifting the partitioning of windows between successive self-attention layers, as depicted in Fig 2.7. These shifted windows establish connections with the windows from the previous layer, substantially boosting the model's power.

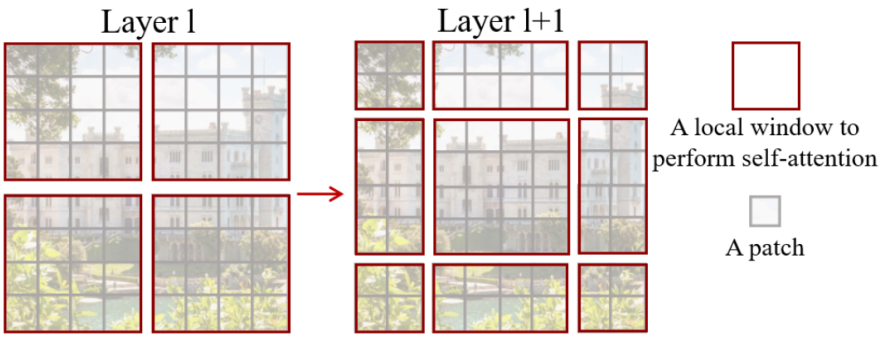

Figure 2.7: Swin Transformer window partitioning for two consecutive layers. [15]

Model Architecture

An overview of the Swin Transformer architecture is illustrated in Fig. 2.8. Similar to the Vision Transformer (see Fig. 2.6), the input image is first split into patches, and their linear embeddings are regarded as tokens. Subsequently, multiple modified self-attention computation blocks, known as Swin Transformer blocks, are applied to these patch tokens. As the network deepens, hierarchical representation is established through patch-merging layers that reduce the number of tokens by combining 2×2 neighboring tokens channel-wise and embedding them into a single token with half channels. Finally, the features from the last layer can be utilized by a classification head for predictions.

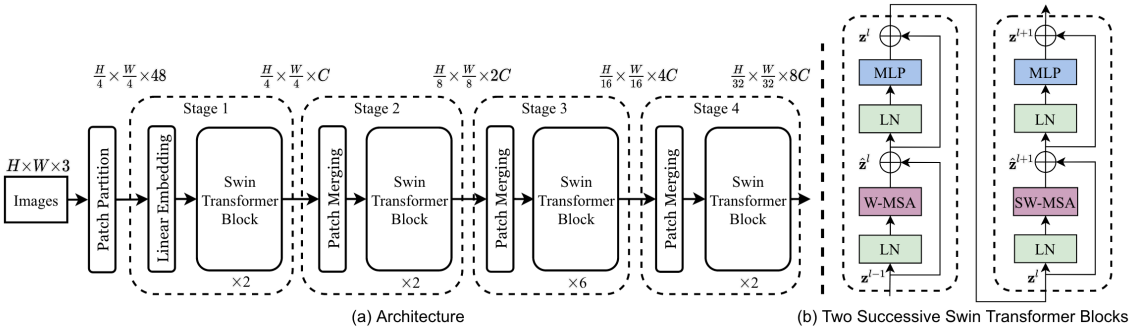

Figure 2.8: Swin Transformer architecture. [15]

In the following, Eq. 11 shows how the output of two successive Swin Transformer blocks (z^{l+1}) is

computed based on their input (z^{l-1}).

$$\hat{z}^l = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1}, \tag{11a}$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l, \tag{11b}$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^l)) + z^l, \tag{11c}$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}, \tag{11d}$$

Here, **W-MSA** and **SW-MSA** are Window MSA and Shifted-Window MSA units whose computations are exactly like Eq. 9, but they are only computed within non-overlapping windows as illustrated in Fig. 2.7.

2.2.4 Focal Modulation Network

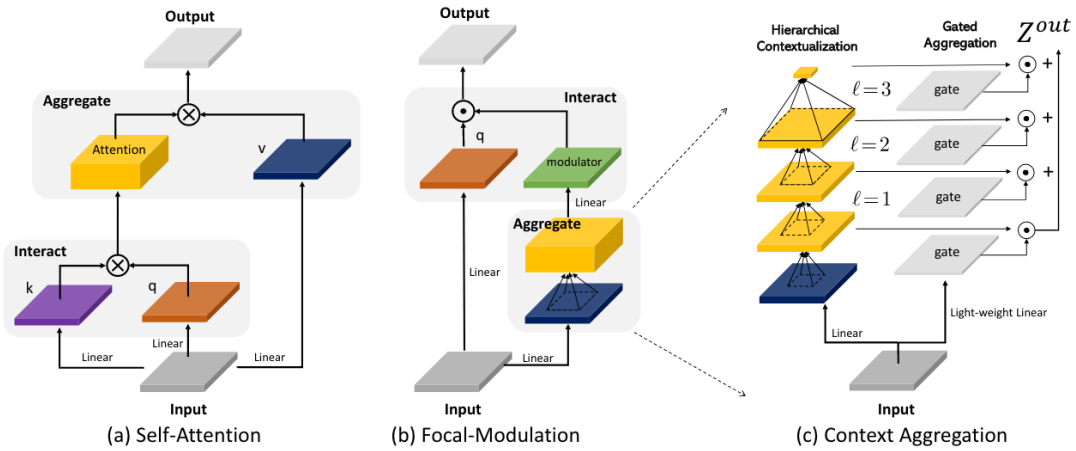

Figure 2.9: Focal Modulation network architecture. [16]

Yang et al. [16] introduced an attention-free mechanism to efficiently compute visual tokens' interactions and dependencies. The proposed Focal Modulation network (FocalNet) resembles the Swin Transformer (refer to Fig. 2.8), where the self-attention units are replaced by Focal modulation blocks. Focal modulation, illustrated in Fig. 2.9, includes hierarchical contextualization using a stack of depth-wise convolutional layers for encoding visual contexts across varying scales, gated

aggregation for context selection based on the query content, and element-wise modulation to integrate aggregated context into the query. In contrast to self-attention, Focal modulation distinctively aggregates contexts at different granularities before modulating individual query tokens, providing an attention-free approach for token interactions [16].

Focal Modulation vs. Self-attention

The key difference between Focal modulation and self-attention lies in the procedural approach: self-attention employs a late aggregation method (see Eq. 12a), wherein aggregation $\mathcal{M}_1$ over the contexts around query token x_i in feature map X occurs after the computation of attention scores between the query and target through interaction $\mathcal{T}_1$. In contrast, Focal modulation employs an early aggregation technique (Eq. 12b), where context features are initially aggregated at each location i using $\mathcal{M}_2$, followed by the interaction of the query with the aggregated feature based on $\mathcal{T}_2$ to yield the refined representation y_i [16].

$$y_i^{\text{self-attention}} = \mathcal{M}_1(\mathcal{T}_1(x_i, X), X), \qquad X \in \mathbb{R}^{H\times W\times C}, x_i \in \mathbb{R}^C, y_i \in \mathbb{R}^C, \tag{12a}$$

$$y_i^{\text{focal modulation}} = \mathcal{T}_2(\mathcal{M}_2(i, X), x_i), \tag{12b}$$

Focal Modulation Operation

The detailed operation of Focal modulation is outlined in Eq. 13:

$$[Q, Z^0, G] = XU_{QZG}, \qquad U_{QZG} \in \mathbb{R}^{C\times(C+C+(L+1))} \tag{13a}$$

$$\{Z^l = f^l(Z^{l-1})\}_{l=1}^{L}, Z^{L+1} = \text{Avg-Pool}(Z^L), \qquad f^l(\mathcal{Z}) \triangleq \text{GeLU}(\text{DW-Conv}(\mathcal{Z}, k^l)) \tag{13b}$$

$$M = U_h\left(\sum_{l=1}^{L+1} G^l \odot Z^l\right), \qquad U_h \in \mathbb{R}^{C\times C} \tag{13c}$$

$$y_i = Q \odot M, \tag{13d}$$

It initiates with linear embedding of the input feature map $X \in \mathbb{R}^{H\times W\times C}$ into Query $Q \in \mathbb{R}^{H\times W\times C}$, feature vector $Z^0 \in \mathbb{R}^{H\times W\times C}$, and Gates $G \in \mathbb{R}^{H\times W\times(L+1)}$ (Eq. 13a). Subsequently, hierarchical contextualization is achieved through a stack of L depth-wise Convolutional layers with

kernel sizes of $k^l = k^{l-1} + 2$ and GeLU activation function. Global average pooling is also applied to the final context granularity Z^L to obtain the global context Z^{L+1} (Eq. 13b). The acquired contexts $\{Z^l\}_{l=1}^{L+1}$ undergo selective aggregation using a Gating mechanism, and the Modulator M is computed by applying the linear layer U_h to the aggregated result (Eq. 13c). Ultimately, the output is attained through element-wise multiplication of the Query and Modulator (Eq. 13d) [16].

2.3 Weakly-supervised Learning

In the domain of medical image segmentation, a significant challenge lies in the acquisition of fully annotated datasets. The process of manually delineating structures or abnormalities in medical images requires expert knowledge and is both time-consuming and resource-intensive. This poses a bottleneck in the development of robust segmentation models, restricting their performance and generalizability to diverse clinical scenarios. Weakly supervised learning emerges as a valuable solution to this challenge by using less detailed labels, like image-level categories, bounding boxes, etc, to train models, skipping the need for precise pixel-wise annotations. This speeds up labeling, allows for larger datasets, and makes medical image segmentation models more effective in diverse real-world situations. This section will explore two prevalent label types utilized in this thesis, bounding boxes and image-level categories, along with the corresponding methods employed for weakly supervised medical image segmentation.

2.3.1 Bounding Boxes

The benefit of using bounding box annotation is its spatial constraints, ensuring that the object of interest is confined within the specified boundaries. This ensures focused learning on relevant regions, potentially leading to better segmentation accuracy. In this section, we delve into two common weakly supervised techniques applied to bounding boxes. The first involves generating pseudo-masks from a shallow segmentation model and refining these initial annotations for precise foreground extraction, with Conditional Random Field (CRF) being a prevalent refinement method. The second approach involves imposing constraints on the network output based on prior knowledge about the target, facilitating end-to-end training.

Conditional Random Field

If we consider the segmentation problem as an energy minimization task, we can define the energy of assigning labeling f to each pixel i as [17]:

$$E(f) = \sum_i \psi_u(f_i) + \sum_{i<j} \psi_p(f_i, f_j) \tag{14a}$$

$$\psi_u(f_i) = -logP(y_i|x, \theta) \tag{14b}$$

$$\psi_p(f_i, f_j) = g(i, j)[f_i \neq f_j] \tag{14c}$$

if a segmentation model with parameters θ generates a label assignment distribution y_i over the input image x, the unary term $\psi_u(f_i)$ is defined as the negative log-likelihood of the probability of this distribution, assessing how well the label f fits at each pixel i. Moreover, the pairwise regularization term $\psi_p(f_i, f_j)$ imposes a penalty on label differences between any two-pixel locations i and j based on the pixel information such as its location and intensity. By minimizing this energy function, the new labels will have a greater likelihood of being assigned to the image.

DeepCut

DeepCut [18] is a weakly supervised DL medical image segmentation technique that leverages bounding boxes to achieve refined tissue segmenation. An illustration of this method is depicted in Fig 2.10. First, we need to emphasize that, on DeepCut, the segmentation is not directly performed on input images. Instead, smaller patches are sampled from the input image, and a Two-Layer CNN classifies these patches into foreground and background. In a naive approach, for training the CNN, any patch inside the initial bounding box is considered the foreground, and patches inside a halo bounding box encompassing the initial one are considered the background. Finally, the coarse segmentation map is further refined using a dense CRF as explained before. However, in DeepCut, this process is performed in multiple iterations, where each iteration further refines the foreground proposals and finetunes the CNN with the new labels. Fig 2.11 shows an example of brain segmentation using DeepCut.

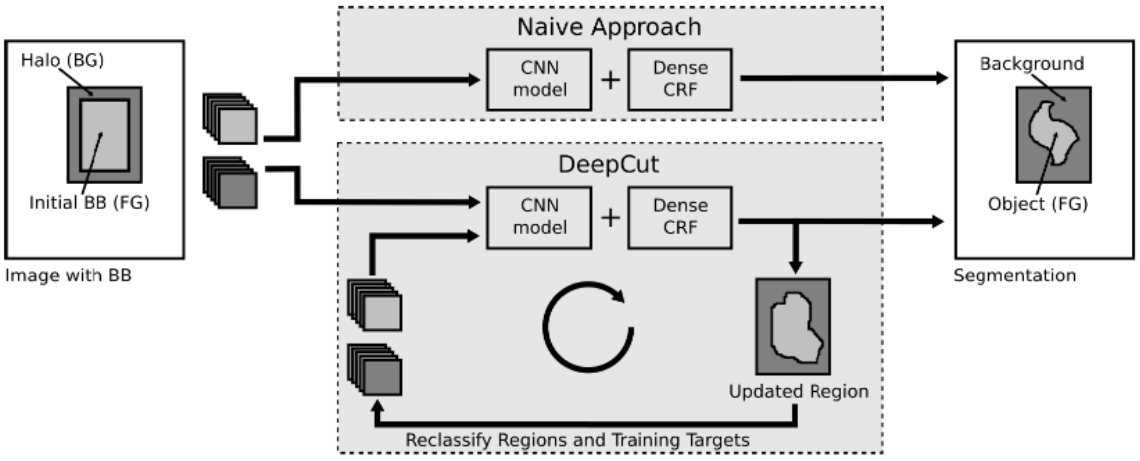

Figure 2.10: DeepCut foreground segmentation vs. Naive CNN segmentation [18]. ©2017 IEEE

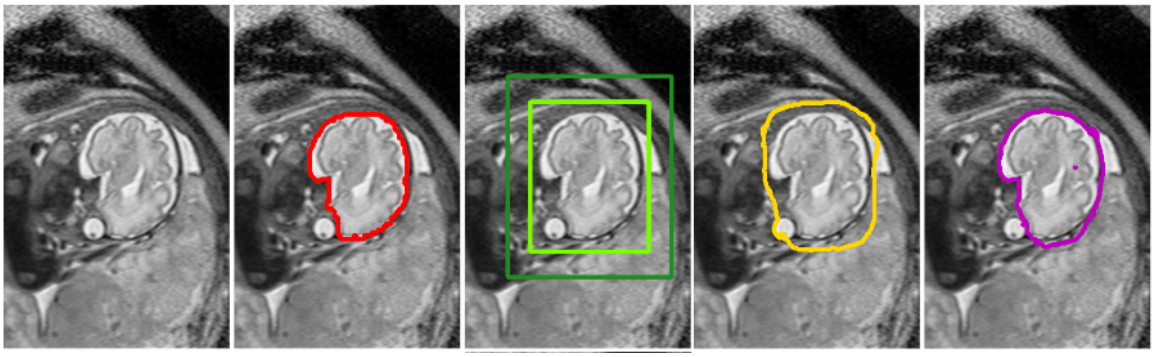

Figure 2.11: DeepCut brain segmentation example: from left to right: (1) original image (2) manual segmentation (red), (3) initial bounding box B with halo H, (4) naive learning approach (one iteration, yellow), (5) DeepCut from bounding boxes (purple) [18]. ©2017 IEEE

Emptiness Constraint and Tightness Prior

Kervadec et al. [19] introduced three constraints on the segmentation network's output based on bounding boxes. The emptiness constraint, accounting for certainty outside the box, stipulates that the predicted foreground size, computed over background pixels Ω_0, should be equal to zero:

$$\sum_{p \in \Omega_0} s_\theta(p) \leq 0, \tag{15}$$

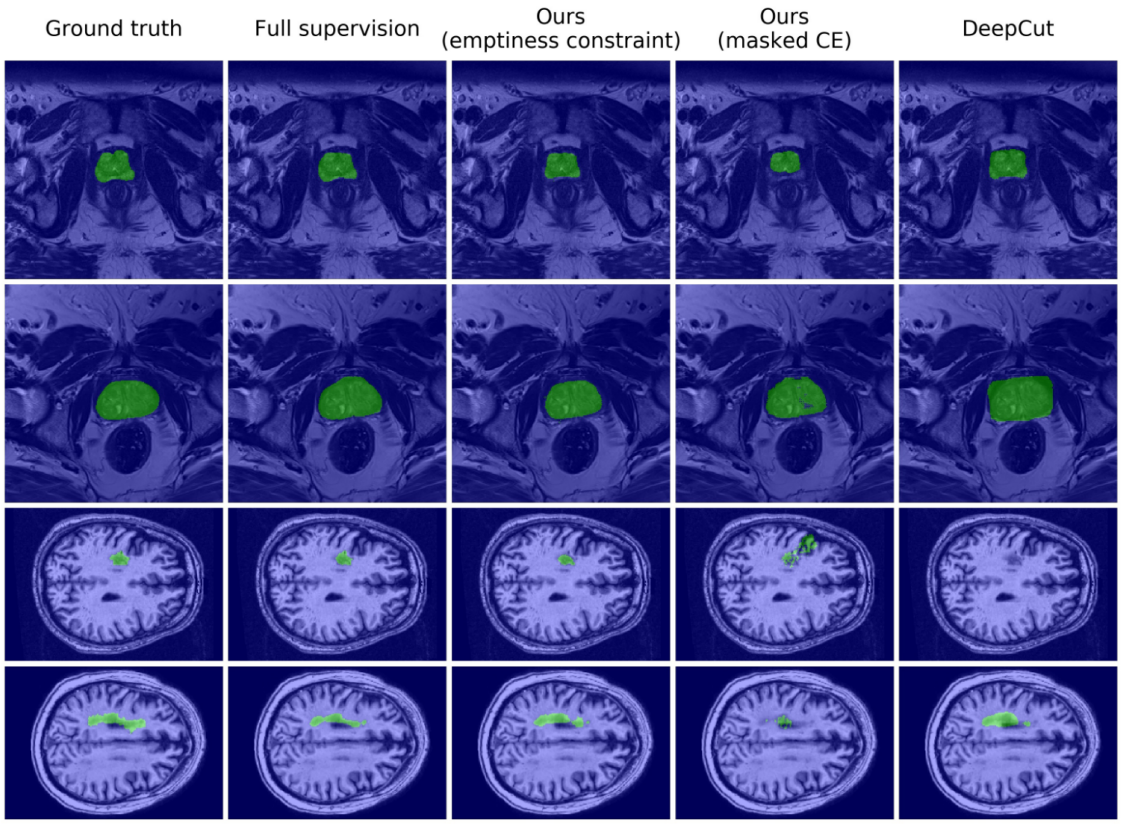

Figure 2.12: A comparison between global constraint approach and DeepCut for weakly supervised segmentation. [19]

Here, the size is defined as the sum of softmax probabilities (s_θ) for ease of computation and differentiability. Conversely, the tightness prior assumes the target region should be sufficiently proximate to each side of the bounding box. It expects each horizontal or vertical line to intersect at least one pixel of the target region, irrespective of its shape. Additionally, regrouping lines into segments of width w, each containing w lines, allows the assumption that at least w pixels of the object will be crossed by the segment:

$$\sum_{p \in s_l} y_p \geq w, \quad \forall s_l \in S_L \tag{16}$$

where $S_L := \{s_l\}$ represents the set of segments parallel to the sides of the bounding boxes. While these constraints might yield trivial solutions — predicting the entire image as background for the emptiness constraint or everything as foreground for the tightness constraint — additional information from the boxes is exploited. The total size of the boxes provides an upper bound on the object

size, and assuming a small fraction ϵ of the box belongs to the target region yields another lower bound:

$$\epsilon|\Omega_I| \leq \sum_{p\in\Omega} s_\theta(p) \leq |\Omega_I| \tag{17}$$

Integrating these global constraints into the network's output, Kervadec et al. [19] conducted a comparative analysis between their weakly supervised approach and the DeepCut model, as illustrated in Fig. 2.12, for prostate segmentation and brain lesion segmentation tasks.

2.3.2 Image Level Categories

Among various weak labels, image-level categories demand the least annotation effort, merely specifying the type of disease without requiring detailed localization like bounding boxes. This simplicity makes it more practical to handle expansive and diverse medical image datasets. In essence, the model is trained to categorize images, and weakly supervised techniques extract segmentation maps from the model's inherent features. In this section, we elaborate on two key approaches for weakly supervised medical image segmentation based on categorical labels, namely class activation maps and attention maps.

Class Activation mapping

Class Activation Mapping (CAM) is a technique demonstrated by Zhou et al. [20] that makes use of the power of Convolutional Neural Networks (CNNs) in localizing the precise regions within an image belonging to specific categories (e.g. Fig. 2.13).

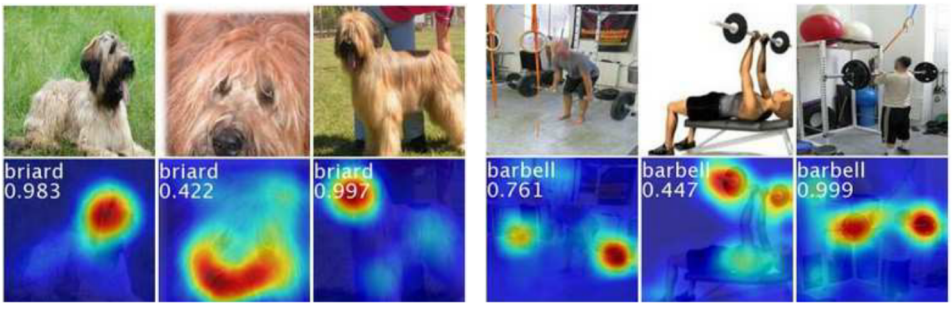

Figure 2.13: An example of Class Activation Maps for two classes. [20]

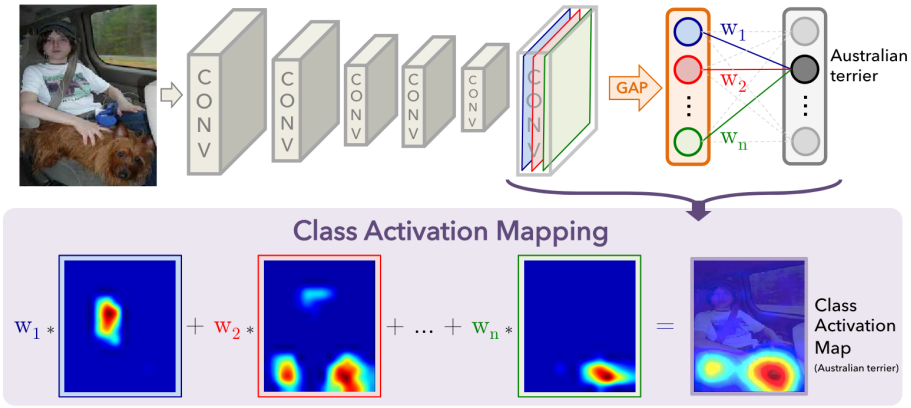

Figure 2.14: Class Activation Mapping technique [20]

As illustrated in Fig. 2.14, the approach involves training an end-to-end architecture primarily comprised of convolutional layers, incorporating global average pooling (GAP) on the final convolutional feature maps. These pooled features then serve as inputs for a fully connected layer responsible for generating the desired classification output. The final convolutional layer consists of $\{f_k\}_{k=1}^{n}$ units, where $f_k(x, y)$ shows the activaton of unit k at spatial location of (x, y). Averaging over all spatials (x, y) yields F_k, the global average of unit k (Eq. 18a). Then, the class c likelihood y^c is computed by softmax(WF) (Eq. 18b, note that the bias is ignored). Each unit k is expected to be triggered by a specific visual pattern within its receptive field, and consequently, f_k serves as a map indicating the presence of this visual pattern. As illustrated in Fig. 2.14 and Eq. 18c, the class activation map $L^c_{CAM} \in \mathbb{R}^{u \times v}$ for any class c is formed by a weighted linear combination of these visual patterns across different spatial locations. The weight between F_k and class c neurons is denoted as w^c_k and it indicates the importance of F_k for class c. Finally, by upsampling the class activation map to the size of the input image, we can identify the image regions most relevant to the particular category.

$$F_k = \frac{1}{Z} \sum_{x,y} f_k(x, y) \tag{18a}$$

$$y^c = \frac{\exp(\sum_k w^c_k F_k)}{\sum_c \exp(\sum_k w^c_k F_k)} \tag{18b}$$

$$L^c_{\text{CAM}}(x, y) = \sum_k w^c_k f_k \tag{18c}$$

Gradient-weighted Class Activation Mapping

The constraint of CAM is its exclusive application to specific CNN architectures that incorporate global average pooling over convolutional maps right before the prediction stage. To overcome this limitation, Selvaraju et al. introduced Gradient-weighted Class Activation Mapping (Grad-CAM) [21]. This method produces visual explanations from any CNN-based network without necessitating architectural modifications or re-training. It achieves this by leveraging the gradients of the target category to weigh the final convolutional feature maps. As depicted in Eq.19 and Fig. 2.15, where $\{f_k\}_{k=1}^{n}$ denotes the k^{th} feature map of the final convolutional layer, the Grad-CAM class discriminative localization map $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ is computed by applying ReLU activation to the linear combination of f_k maps, weighted by the average gradient of the class score y^c concerning these maps. The ReLU activation is employed to selectively retain features with a *positive* impact on the target class, while ignoring the inclusion of negative values that may pertain to other categories.

$$\alpha_k^c = \frac{1}{Z} \sum_{x,y} \frac{\partial y^c}{\partial f_k(x,y)} \tag{19a}$$

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c f_k \right) \tag{19b}$$

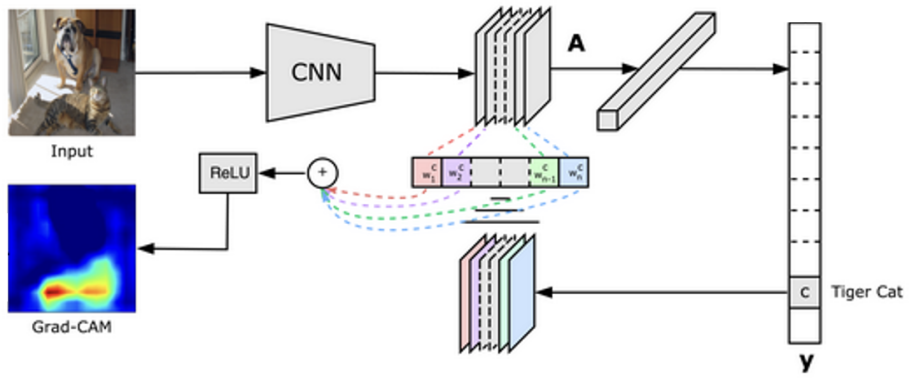

Figure 2.15: Gradient-weighted Class Activation Mapping technique [21]

Pneumothorax Segmentation on Chest X-Ray Images

Leveraging the Grad-CAM approach, Viniavskyi et al. [22] achieved 0.7677 mean IoU for weakly supervised segmentation of pneumothorax X-ray images using a three-step methodology:

(1) **Class Activation Map generation:** First, a fully supervised classification model is trained based on image-level labels. This trained model is then employed to generate activation maps using the Grad-CAM method, and the derived maps function as pseudo labels for the segmentation task.

(2) **Refinement:** In the next step, the acquired pseudo labels undergo refinement via thresholding, dense CRF post-processing, and application of the Inter-pixel Relations network (IRNet) [23]. IRNet leverages both background and foreground class activation maps to effectively enhance the boundaries between them.

(3) **Segmentation:** In the last step, a fully supervised segmentation model, such as U-Net, is trained using the maps generated in the preceding step to perform segmentation.

Fig. 2.16 illustrates an example of these three steps applied an an X-ray image.

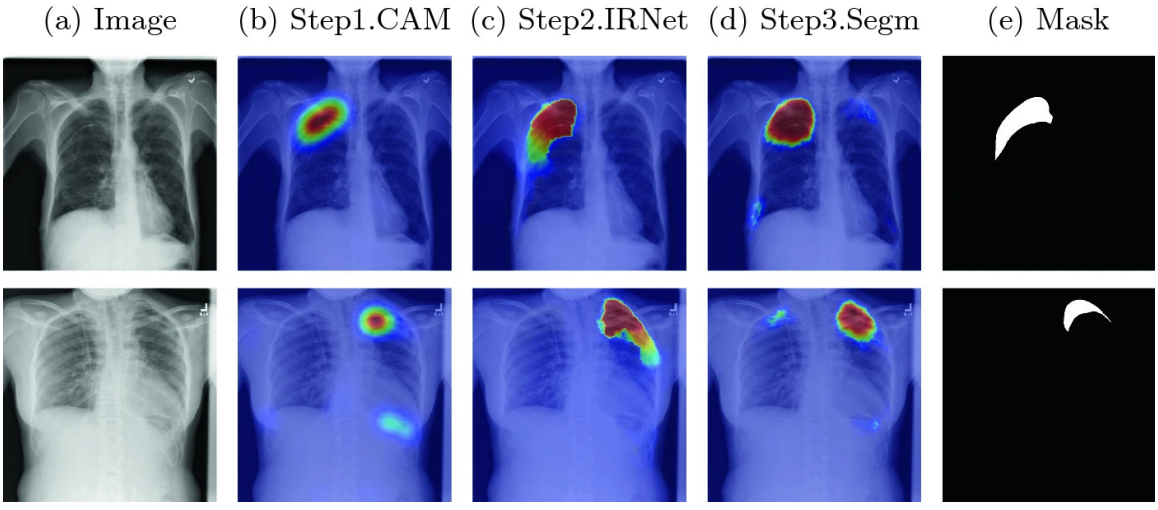

Figure 2.16: Example of weakly supervised pneumothorax segmentation using CAM, IRNet, and U-Net [22].

Intracranial Hemorrhage Detection Explainability

Salehinejad et al. [24] trained a fully supervised classification model on CT images, targeting five categories of distinct Intracranial Hemorrhage (ICH) subtypes: Intraventricular (IVH), Intraparenchymal (IPH), Subarachnoid (SAH), Epidural (EDH), and Subdural (SDH). Subsequently, to illustrate the interpretability of this network, they visualized Grad-CAM maps for various hemorrhages, as depicted in Fig. 2.17, and they confirmed these maps with a radiologist.

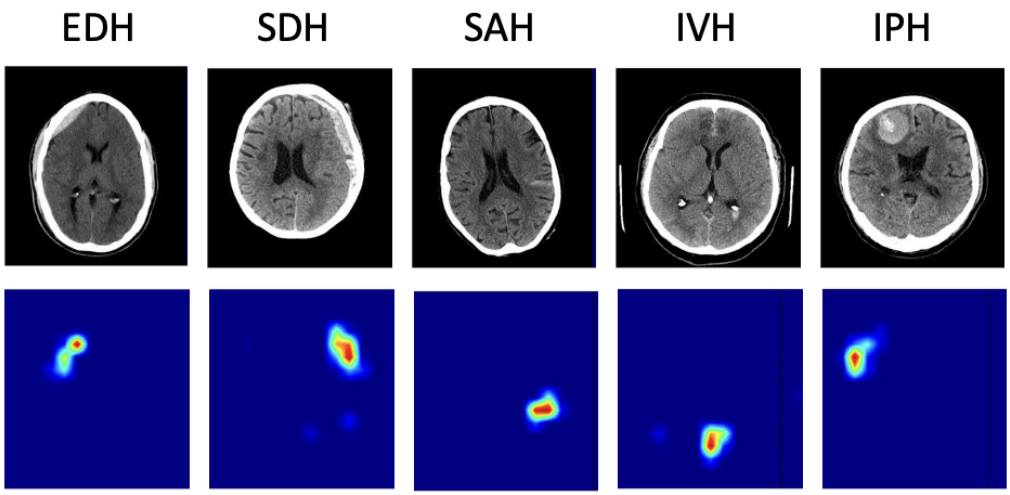

Figure 2.17: Example of Grad-CAM maps derived from an ICH detection model [24].

Attention map

In an attention-based model such as ViT [14], attention maps provide interpretability by unveiling the precise regions within an input image that the model deems critical for its decision-making process. Figure 2.18 showcases eight examples where the attention map of the Vision Transformer is superimposed onto the input images.

In an L layer Vision Transformer, the attention weight of layer l is defined as $A_l \in \mathbb{R}^{N\times N}$ (see Eq. 9), where N is the number of tokens (refer to Sec. 2.2.2), and the entry $A_{i,j}$ signifies the similarity between tokens i and j. Defining $\hat{A}$ as the matrix multiplication of attention weights across all layers ($\hat{A} = \Pi_{l=1}^{L} A_l$), the resulting attention map is denoted as $\hat{A}_0$, representing the weight of the class-token. To the best of our knowledge, the utilization of self-attention maps

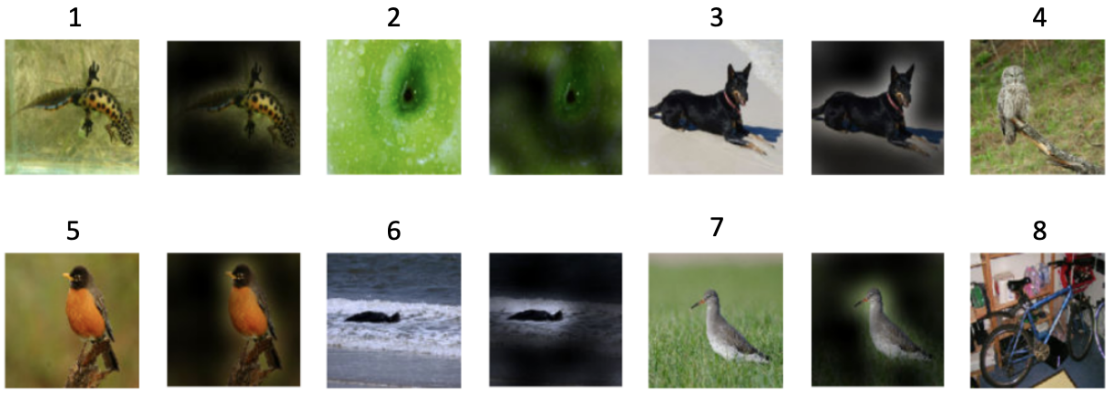

Figure 2.18: Visualization of Vision Transformer Attention map [14]

derived from Transformer-based models for weakly supervised medical image segmentation has not been investigated before our study in Chapter 3, motivating us to explore it in this thesis.

Chapter 3

Intracranial Hemorrhage Segmentation Using Hierarchical Combination of Attention Maps

A version of this chapter was presented at the MLCN (Machine Learning for Clinical Neuroimaging) 2022 joint workshop at the Medical Image Computing and Computer Assisted Interventions (MICCAI) Conference.

- **A. Rasoulian**, S. Salari, and Y. Xiao, "Weakly supervised intracranial hemorrhage segmentation using hierarchical combination of attention maps from a swin transformer," in Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings, pp. 63–72, Springer, 2022.

3.1 Introduction

Intracranial Hemorrhage (ICH) is the most deadly type of cerebrovascular disease, accounting for 10-15% of all stroke cases [25, 26]. The outcome is highly correlated with the hemorrhage volume, which is susceptible to enlarge in the first three hours [27]. Thus, there is a high risk for ICH to

turn into a secondary brain injury or even death if it is not treated in time. Depending on the location of hemorrhage in the brain, ICH can be divided into five subtypes: Intraventricular (IVH), Intraparenchymal (IPH), Subarachnoid (SAH), Epidural (EDH), and Subdural (SDH). Treatment methods must be tailored towards specific ICH subtypes, and a surgery is done only if the location of hemorrhage is favorable. Rapid and accurate detection and quantification of ICH is therefore crucial in choosing correct treatments and thus reduction of patient mortality. With quick imaging time and good accessibility, computerized tomography (CT) is commonly used in the clinic to assess ICH.

Previous developments in convolutional neural networks (CNNs) have resulted in a great number of fast and accurate solutions in computer-assisted diagnosis and treatment decisions, in the forms of image classification and/or segmentation, including those for the care of ICH [28]. One issue with the CNNs is their limited capacity to encode long-range spatial information, but it may affect ICH detection/subtyping accuracy as the location of hemorrhage is directly relevant to the diagnosis. Recently, Dosovitskiy et al.[14] introduced the Vision Transformer (ViT), which has attracted significant interests for vision tasks, especially in the context of medical imaging [29, 30], where multi-head attention mechanisms are used to encode the contextual relationship between image patches (as tokens). However, compared with CNNs, the ViT has low locality inductive biases (e.g., translational invariant features). As a recent variant to mitigate the drawback of the ViT, the Swin transformer [15] is an efficient hierarchical transformer that gradually reduces the number of tokens by merging image patches and computing attentions in non-overlapping local windows. To the best of our knowledge, the Swin transformer has not been used for ICH detection or segmentation. For CNNs and especially transformer-based models, a large amount of training data is necessary. However, annotating medical images is laborious and time-consuming, especially for segmentation tasks. One way to mitigate this problem is through weak supervision [31], where more accessible or coarse annotations (e.g., categorical labels or bounding boxes) are used to generate more refined ones, such as segmentation masks.

In this work, we built a novel weakly supervised framework for ICH segmentation leveraging the attention maps generated from a Swin transformer, which is trained using categorical labels for ICH detection based on public databases. As an exploratory investigation, our study has three major

contributions. **First**, the Swin transformer is used for ICH detection for the first time. **Second**, we proposed a new method to obtain ICH segmentation by leveraging the hierarchical combination of self-attention maps from the trained ICH detection transformer, and demonstrated its feasibility and performance. **Lastly**, to examine the impact of learning tasks on self-attention maps and weakly supervised segmentation, we compared the segmentation performance for two Swin transformers based on (1) binary classification (presence of hemorrhage or not) and (2) multi-label classification (detailed ICH subtypes and with/without ICH).

3.2 Related Works

Several techniques have been proposed for the detection and segmentation of ICH. An excellent recent review is provided by Hssayeni et al. [28], with almost all using supervised learning strategies in semi-automatic and automatic manners, achieving the area-under-the-curve (AUC) of 0.846∼0.975 for binary classification (ICH vs. without ICH) and 0.93∼0.96 for ICH subtyping. For deep learning-based approaches, fully convolutional networks (FCNs) [32] and recurrent neural networks (RNNs) [33] models were often used, and the accuracy in ICH vs. without ICH classification was shown to be higher than ICH sub-typing in general [28]. With interests in explainable CNNs, attention mechanisms have been deployed to enhance detection accuracy and visualize the region of interest for the classification results [24]. The latter also inspired their application for weakly supervised brain lesion/hemorrhage segmentation, which has been attempted by only a few [34, 35]. Earlier, Wu et al. [34] employed refined 3D Class-Activation Maps (CAMs) to learn a representation model for brain lesion segmentation and achieved a 0.3827 mean Dice score on the Ischemic Stroke Lesion Segmentation (ISLES) dataset (multi-spectral MRI). Similarly, Nemcek et al. [35] detected the location of ICH as bounding boxes in axial brain CT slices using the local extrema of attention maps obtained from a ResNet-like binary classification CNN, and they achieved a mean Dice of 0.58 for the lesion bounding boxes. So far, self-attention has not been experimented for weakly supervised ICH segmentation, thus motivating us to explore it in this study.

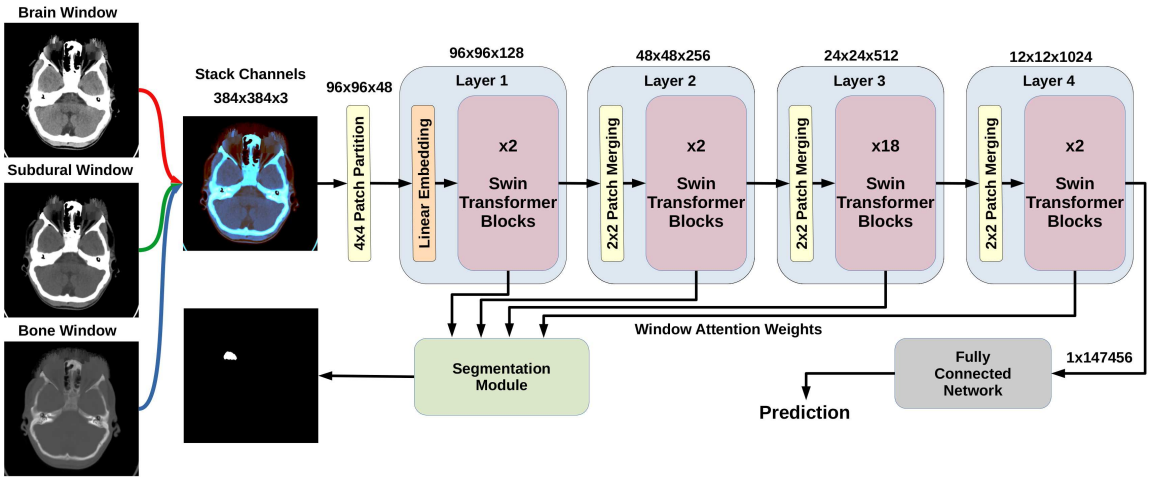

Figure 3.1: An overview of the proposed weakly supervised ICH segmentation algorithm

3.3 Proposed methods

An overview of the proposed weakly supervised segmentation technique is depicted in Fig. 3.1, where it is divided into three components. First, we train a deep learning (DL) model with a Swin transformer as the backbone for categorical classification of ICH. Then, during test time, we obtain hierarchical layer-wise attention maps for the input image from the trained model. Finally, segmentation is achieved by binarizing the hemorrhage localization map made by combining the window attention maps and soft tissue intensity information. Note that one patient may have multiple ICH subtypes. Since the CT data were from several clinical centers with different slice thicknesses, we decided to implement our algorithm for 2D axial CT slices.

ICH classification:

For the proposed technique, we used the Swin-B transformer pretrained and finetuned on ImageNet1K and ImageNet21K datasets [36]. Each two successive Swin transformer blocks have window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) units for computing attention weights (see Fig.2a) [15]. Here, the shifted windowing scheme helps establish connections between windows, in comparison to the ViT. To investigate the impact of different arrangments of categorical learning on the self-attention maps and thus the proposed weakly

supervised segmentation, we trained two versions of the Swin transformer model for 1) binary classification (ICH vs. without ICH) and 2) multi-label classification (recognizing 5 ICH subtypes and with/without ICH). To address the issue of the imbalanced dataset, we use the focal binary cross-entropy [37] loss function to train our model:

$$loss = \frac{1}{N}\sum_{k=1}^{N} Y_k \cdot (1-y_k)^{\gamma} \cdot log(y_k) + (1-Y_k) \cdot y_k^{\gamma} \cdot log(1-y_k) \tag{20}$$

Here N, Y, y, and γ are batch size, ground-truth label, sigmoid of predicted output, and the focal loss focusing parameter, respectively. As Lin et al. [37] suggested, we set the value of γ=2. For multi-label classification, the overall loss is the weighted average of subtypes' losses computed above, where each of five subtypes' weight is 1, and ICH vs. without ICH weight is 2.

Attention map generation:

For our technique, we decided to employ the raw attention weights of all layers to obtain the relevant attention maps for weakly supervised segmentation, instead of the more commonly used visualization of class activation mapping (CAM). This is due to two reasons. First, we would like to fully leverage the relevant information from earlier layers considering the Swin transformer architecture. Second, without gradient computation, the processing can be more efficient.

In previous attempts to visualize attention weights with the ViT, an additional classification token was added to the image patches, and after multiplying the attention weights of all layers, only this token's weight was retrieved as the attention map [38, 14]. However, as the Swin transformer uses a windowing method, adding another token corrupts the window division. Besides, attention weights at every two successive MSA units correspond to regular and shifted image patches, and multiplying them is meaningless. Hence, instead of multiplying weights, we compute the attention map at each unit, and then we multiply their respective maps. Here, Fig. 3.2 shows the steps of producing the layer-wise attention map from two successive Swin transformer blocks. First, we perform Global Average Pooling (GAP) on all tokens' attention weights; Then, a full-image map is reproduced by concatenating window-wise maps. Note that an additional step of reverse shifting is needed for SW-MSA units, and the results from W-MSA and SW-MSA are multiplied to produce the layer attention

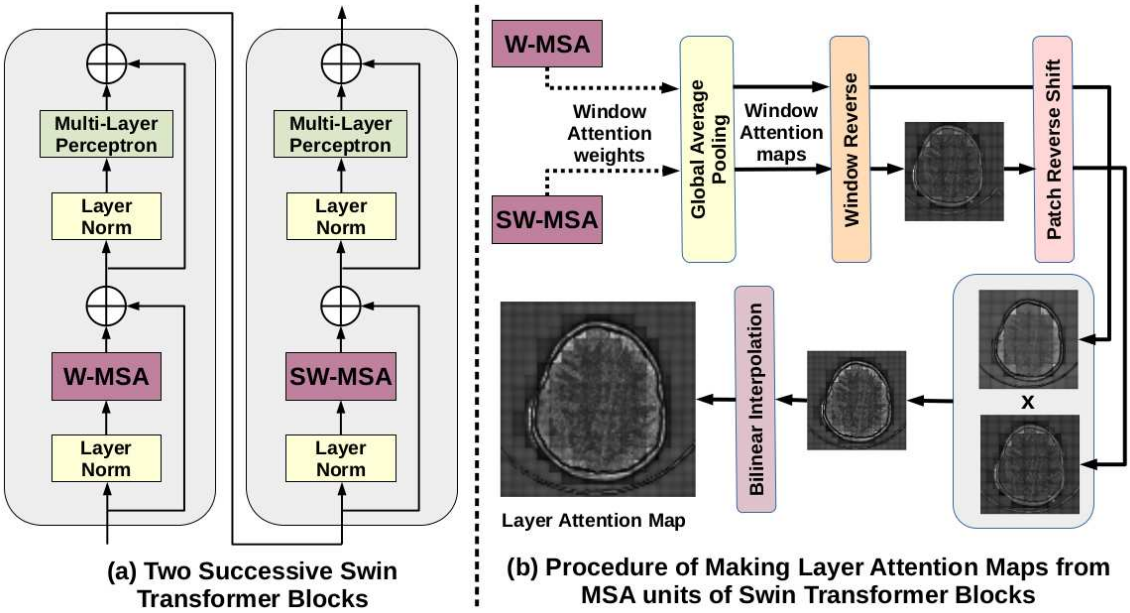

Figure 3.2: Generation of layer-wise attention maps from the Swin transformer

map. Inspired by [39, 40, 41], we finally combine different layers' attention maps at the resolution of the input image, and bilinear interpolation is used when matching the resolutions across the maps at different layers/hierarchies. This technique helps produce more precise attention visualization for segmentation purposes. Lastly, the hemorrhage localization map (see Fig. 3.3) is produced by multiplying the resulting self-attention map with the "brain-tissue window" channel from the CT slice to enhance the discriminative power for ICH identification.

Discrete segmentation:

The final discrete ICH segmentation is obtained by binarizing the hemorrhage localization map (see Fig. 3.3). We experimented with three binarization techniques, including simple thresholding, Otsu's method, and k-means, and selected simple thresholding as the optimal choice due to its performance. For simple thresholding, the threshold value is computed as $t = S \times M_{max}$ where M_{max} is the maximum intensity in the hemorrhage localization map, and S is a scalar. We used 10-fold cross-validation on the test data to find an appropriate value for S between 0 to 1 with a step size of 0.01.

To compare with the proposed technique, we also implemented a similar weakly supervised ICH

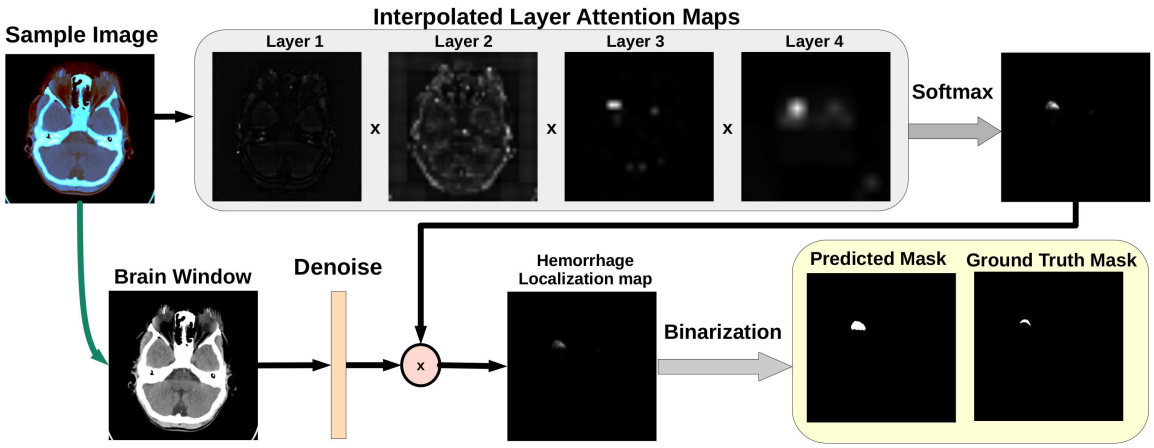

Figure 3.3: Procedure for generating ICH segmentation based on the combination of hierarchical attention maps from the ICH detection transformer.

segmentation method based on binary ICH classification with the Swin transformer and GradCAM [21], which was implemented in two versions: one only on the last layer and the other with a similar hierarchical approach to obtain the attention maps.

3.4 Experiments and Results

3.4.1 Dataset

To implement and validate our proposed algorithm, we employed the public RSNA 2019 ICH [42] and PhysioNet[28] CT datasets. The RNSA dataset contains 752,803 CT slices, with each slice annotated with ICH subtypes. On the other hand, the PhysioNet dataset has 2,814 CT slices, and ICH was manually segmented while ICH subtypes are also provided. Only the RSNA dataset was used for training, and we used the PhysioNet dataset as a separate testing set to examine the performance of our model. For each CT slice, brain, subdural and bone windows created using the suggested parameters provided in the relevant data publications [42, 28] were stacked to create a three-channel image and are downsampled to 384×384 pixels.

3.4.2 Imlpementation and evaluation

The training dataset (RSNA2019 ICH) was randomly split into 90% and 10% for the training and validation sets. We employed the AdamW optimizer with an initial learning rate of 1e-5 for model

training. In addition, an early stopping with patience=3 was used to stop training if the validation loss did not decrease for three consecutive epochs. To improve the robustness of our model, data augmentation techniques including random left-right flipping, image rotation, and Gaussian noise addition were also used. The focal binary cross-entropy was used as the loss function to tackle the imbalanced data in the training dataset, where much more CT slices without ICH exist. The network was trained on a desktop computer with an Intel Core i9 CPU and a NVIDIA GeForce RTX 3090 GPU with 24GB memory. To test the performance of the ICH detection accuracy and the performance of hemorrhage segmentation, the PhysioNet-ICH data was used. The accuracy, AUC, specificity, and F1-score were evaluated for the classification tasks, and for segmentation, the Dice coefficient is reported. When assessing the multi-label classification model against the binary classification one, an image is categorized as ICH if any subtypes are detected. Thus, the differences in ICH detection between the two DL models were confirmed using a chi-square test, and the Dice coefficients for segmentation performance between the two models were compared using a two-sided paired-sample t-test.

3.4.3 Results

The results of our experiments are listed in Table 3.1 for the automatic detection and weakly supervised segmentation of ICH when employing binary and multi-label classification tasks. In terms of the quality of ICH detection, there is no significant difference between the two proposed DL models for ICH vs. without ICH classification ($p>0.05$), while the binary classification achieves an AUC of 0.974. Regarding ICH subtyping, we have achieved the AUCs of 0.941, 0.976, 0.996, 0.965, and 0.984 for EDH, IPH, IVH, SAH, and SDH, respectively. For hemorrhage segmentation, the binary classification model yielded a mean Dice of 0.407 (with simple thresholding), which is significantly higher than the multi-label counterpart ($p<0.05$). The same trend holds for the other two image binarization methods ($p<0.05$). When comparing different binarization methods, simple thresholding offers the best results, potentially due to the high imbalance between ICH and non-ICH pixels. Furthermore, as an ablation study, the GradCAM-based methods, when applied to the final layer and with a similar hierarchical approach, could achieve 0.187 and 0.100 mean Dice scores (also using the simple thresholding method), respectively, which are far worse than our proposed

Table 3.1: ICH detection and weakly supervised segmentation results for binary and multi-label ICH classification models (reported values for Dice are mean±std)

	ICH detection				ICH segmentation - Dice		
	Accuracy	**AUC**	**Specificity**	**F1-score**	**Simple thresholding**	**Otsu's method**	**k-means**
Binary	0.953	0.974	0.973	0.791	0.407±0.225	0.383±0.228	0.326±0.228
Multi-label	0.952	0.975	0.979	0.776	0.324±0.237	0.316±0.246	0.268±0.229

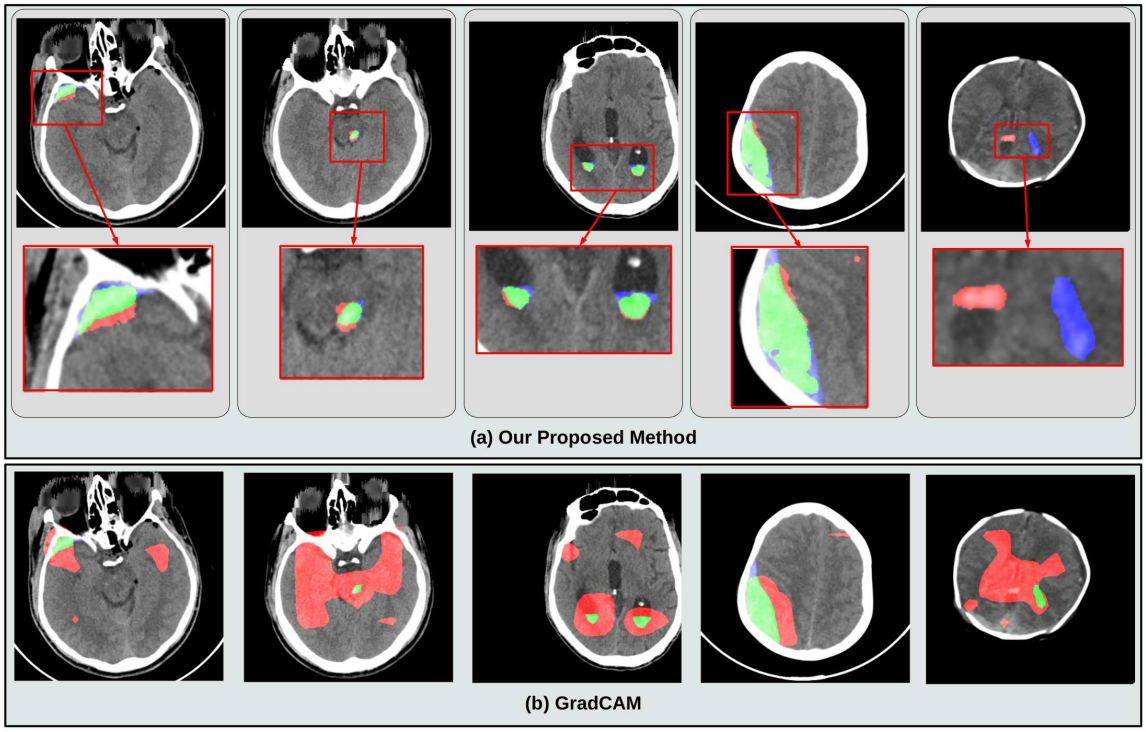

Figure 3.4: Demonstration of weakly supervised ICH segmentation with five cases with close-up views. Green=True Positive, Blue=False Negative, Red=False Positive.

method. Finally, a qualitative demonstration of the segmentation results is shown in Fig. 3.4 for five different cases (each case per column) between the proposed method and the GradCAM approach. With the visual demonstrations, we can see that the final segmentation with the proposed method produces better results than the GradCAM approach, which provides good coverage for the hemorrhage regions, but often overestimates the extent.

3.5 Discussion

As the first attempt to use the Swin transformer for ICH detection, we obtained an accuracy of AUC=0.974. For both cases of binary and multi-label classifications, these results are in line with or better than previous reports [28]. Since there were very few studies on weakly-supervised pixel/voxel-wise ICH segmentation, it is difficult to assess our method against the state-of-the-art. The closest prior work is the technique by [34], which was employed to segment stroke lesions from multi-spectral MRI with a Dice score of 0.3827. For ICH segmentation, a similar but potentially more challenging task due to the small size, irregular shape, and subtle contrast of the target in CT, our technique has achieved a higher Dice score (0.407). As an additional reference, in the original data paper of PhysioNet[28], a supervised U-Net achieved a Dice of 0.315 for ICH segmentation. Although the overall performance of weakly supervised brain lesion/hemorrhage segmentation is still inferior to the supervised counterparts, the relevant explorations, like the presented study are valuable in mitigating the heavy reliance on detailed image annotations.

The attention mechanism has been popular to improve the transparency of deep learning algorithms and has been the focus of many weakly supervised segmentation algorithms [34, 35]. Nevertheless, the impact of different learning strategies on attention maps was rarely investigated. In this exploratory study, we examined such an impact through the example of ICH detection by comparing binary and multi-label classifications. As our weakly-supervised ICH segmentation heavily relies on the resulting self-attention maps, the segmentation accuracy also reflects how well the network focuses on the relevant regions for the designated diagnostic task. Based on our observations using the Swin transformer, binary classification offers better overlap between the network's attention and the relevant pathological region, while both strategies offer similar performance for ICH detection (grouping all subtypes as ICH). Future exploration is still needed to better understand the observed trend.

Several aspects can still be explored to further improve our segmentation performance in the future. First, instead of 2D slice processing, inter-slice or 3D spatial information may be incorporated to enhance the performance of ICH detection and segmentation. Second, more efficient and elaborate learning-based methods can be devised to further refine the initial segmentation obtained

with self-attention maps to allow better segmentation accuracy.

3.6 Conclusion

In conclusion, leveraging the Swin transformer and public datasets, we have developed a framework for weakly supervised segmentation of ICH based on categorical labels. To tackle the issue of limited and expensive training data for ICH segmentation, we have showcased the feasibility of this approach and further demonstrated the benefit of binary classification over multi-label classification in weakly supervised segmentation. With these insights, future studies could further improve the proposed technique's accuracy and robustness.

Chapter 4

Intracranial Hemorrhage Segmentation Using Head-wise Gradient-infused Self-attention Maps

A version of this chapter was published as an invited paper in the Machine Learning for Biomedical Imaging (MELBA) journal.

- **A. Rasoulian**, S. Salari, and Y. Xiao, "Weakly supervised intracranial hemorrhage segmentation using head-wise gradient-infused self-attention maps from a swin transformer in categorical learning," Machine Learning for Biomedical Imaging, vol. 2, pp. 338–360, 2023.

4.1 Introduction

Intracranial Hemorrhage (ICH) is a potentially fatal cerebrovascular disorder that is responsible for 10-15% of all stroke cases and can be caused by various factors, such as head trauma, high blood pressure, and blood clots [25, 26]. The outcome of ICH depends on the volume of bleeding, which can enlarge rapidly within the first few hours [27], leading to a high risk of secondary brain injury or even death if it is not treated promptly. In general, ICH can be classified into five subtypes based on its location in the brain, including Intraventricular (IVH), Intraparenchymal (IPH), Subarachnoid

(SAH), Epidural (EDH), and Subdural (SDH). Note that one patient may have more than one hemorrhage subtype. Each ICH subtype should receive customized treatment approaches, and surgery is only considered if the location of the hemorrhage is advantageous. Upon admission at the hospital, early detection and accurate quantification of ICH are critical in selecting appropriate medical interventions and reducing patient mortality. Thus, efficient and automated systems to asses ICH are highly valuable. Compared to other medical imaging modalities, such as MRI, computerized tomography (CT) is often used in the clinic to assess ICH due to its fast imaging time and good accessibility. However, in addition to the morphological and spatial variabilities, the subtle contrast of ICH within often noisy clinical CT scans can pose challenges in its detection and quantification.

Recent progresses in deep learning (DL) techniques, especially convolutional neural networks (CNNs), have led to the development of efficient and accurate solutions for computer-assisted diagnosis and treatment decisions. For the care of intracranial hemorrhage, several automatic CNN-based DL algorithms have been devised for the detection, subtyping, and volumetric segmentation of intracranial hemorrhage based on clinical scans [28, 43, 24]. To overcome the limitations of CNNs in encoding long-range spatial information due to limited field of view, which may impact the accuracy of ICH detection and subtyping, particularly in cases where the spatial location of hemorrhage is crucial for diagnosis, the Vision Transformer (ViT) [14] has emerged as a promising solution. The ViT utilizes multi-head attention mechanisms to capture contextual relationships among spatially distributed image patches and has attracted great interest for vision tasks, including medical imaging applications [29, 30]. However, by removing convolutions, the ViT possesses low locality inductive biases, such as translation invariant features. To address this, a recent variant called the Swin transformer [15] was introduced as an efficient hierarchical transformer, addressing the need for both long-range spatial encoding and local feature representation. It achieves the goal by gradually reducing the number of tokens by merging image patches and computing attention in non-overlapping local windows to mitigate the drawback of the ViT.

Training CNNs and Transformer-based models require a significant amount of data, but annotating medical images is a laborious and time-consuming process, particularly for segmentation tasks. Among various strategies, including semi-supervised learning, weakly supervised methods

[31] offer alternative solutions to address such challenges by deriving fine-grained image segmentation from coarse and more accessible image annotations, such as bounding boxes, scribbles, and categorical labels. Among these typical choices, as categorical labels require the least time and effort, obtaining pixel-wise segmentation from them is highly attractive. This is especially true for our target application, where image classification is also needed, but such approaches have rarely been attempted. In this study, we intend to propose and validate a novel weakly supervised ICH segmentation technique by taking advantage of the Swin transformer.

In our previous work [44], we employed a Swin transformer to perform CT-based detection and weakly supervised segmentation of ICH for the first time. More specifically, we obtained ICH segmentation by fusing hierarchical self-attention maps generated from a Swin transformer that was trained using categorical labels for ICH detection. Furthermore, comparing the proposed weakly supervised ICH segmentation framework for two Swin transformers based on (1) binary classification (presence of hemorrhage or not) and (2) multi-label classification (detailed ICH subtypes and with/without ICH), we found that binary classification helped better focus the network attention on the ICH regions. In this paper, we further extended our previous study [44] with three main contributions. **First**, inspired by the gradient-weighted class activation mapping (Grad-CAM) [21], we proposed a novel attention visualization technique, called HGI-SAM (Head-wise Gradient-infused Self-Attention Mapping), by performing head-wise weighing of self-attention obtained from the Swin transformer using the gradient of the target class. We further demonstrated the benefit of incorporating HGI-SAM in our weakly supervised ICH segmentation framework over the original proposal [44]. **Second**, by inspecting the characteristics of the gradient-weighted attention maps obtained from ICH detection, we proposed tailored post-processing methods to optimize the segmentation accuracy. **Lastly**, with the publicly available RSNA 2019 Brain CT hemorrhage [42] and PhysioNet datasets [28], we conducted a comprehensive evaluation of the new method against our previous approaches, popular U-Net and Swin-UNETR models with full supervision, and a similar weakly supervised segmentation method leveraging the popular Grad-CAM technique, in the tasks of ICH segmentation and detection.

4.2 Related Works

There have been several variants of the Swin transformer model for medical image segmentation tasks. Heidari et al. [45] introduced a model with an encoder that combines feature maps from a CNN and a Swin transformer, to achieve accurate segmentation of skin lesions, multiple myeloma cells, and abdominal CT scans. Cao et al. [46] proposed a Swin-based U-Net-like model to segment abdominal CT images and cardiac MRI scans. Hatamizadeh et al. [47] proposed a hybrid Swin-encoder-CNN-decoder model to segment brain tumor MRI images. Finally, Lin et al. [48] introduced a dual Swin transformer model with different patch sizes to segment endoscopic images. Although all these methods showcase promising results to demonstrate the capability of the Swin transformer architecture, they all require full supervision.

To overcome the challenge of limited, well-annotated training data in developing deep learning techniques for medical image segmentation, a number of semi-supervised and weakly supervised algorithms have been proposed [49, 50, 51]. Semi-supervised strategies leverage a small number of images with refined labels, along with unlabeled or weakly labeled data. In this domain, Yurt et al. [52] used Generative Adversarial Networks (GANs) for MRI contrast translation with undersampled k-space data. Chen et al. [53] employed attention-based multi-task learning that simultaneously optimizes a supervised segmentation and an unsupervised reconstruction for brain tumor segmentation. Finally, Zhou et al. [54] incorporated collaborative learning for diabetic retinopathy grading and lesion segmentation. On the other hand, weakly supervised techniques rely entirely on coarse labels in the formats of bounding boxes [18], scribbles [55], points [56], or even categorical labels [57]. As these coarse-level labels are more economical to acquire, weakly supervised segmentation techniques can further reduce the need for refined pixel/voxel-level annotations. With simple bounding boxes, Rajchl et al. [18] proposed DeepCut, an approach that combined a CNN segmentation model with a densely-connected conditional random field (CRF) in an iterative training process to achieve pixel-level segmentation. Their method was tested on brain and lung segmentation for fetal MRI datasets. Following the approach, Kervadec et al. [19] employed global constraints derived from box annotations, including tightness prior and global background emptiness, to achieve improved segmentation results over DeepCut [18] on the PROMISE12 dataset [58]. Previously, scribble and

point annotations have been widely used in interactive segmentation. In weakly supervised segmentation, Roth et al. [56] employed the random walker algorithm to generate coarse image-level labels from anatomical landmarks, which were used in combination with the point clouds to refine the segmentation results. More recently, Liu et al. [55] proposed a weakly supervised COVID-19 infection segmentation method based on image scribbles and an uncertainty-aware mean teacher framework.

To further alleviate the need for pixel/voxel-wise manual annotation, weakly supervised segmentation methods that solely rely on categorical labels are highly attractive. With the assumption that deep neural networks in image classification tasks should have a local focus on the target objects, this type of approach was made possible by the latest techniques that provide an intuitive visual explanation of the reasoning process for DL algorithms through saliency, class activation, and attention maps. In this domain, Han et al. [59] proposed a weakly supervised segmentation model based on class residual attention for the lung adenocarcinoma and breast cancer datasets. Chen et al. [60] developed a novel class activation mapping for weakly supervised segmentation for MRI datasets that achieves state-of-the-art accuracy, and similarly, Viniavskyi et al. [22] utilized class activation maps (CAM) for Chest X-Ray segmentation. More recently, Yu et al. [61] further modified CAMs by scale feature adaptation and soft-erase modules to segment thyroid ultrasound images. With the transformer model, Li et al. [62] utilized a self-attention mechanism in multiple instances learning for weakly supervised segmentation of histopathology images while Zhang et al. [63] used CAM and a refinement segmentation decoder for the same task.

Almost all previous reports on automatic ICH detection and/or segmentation primarily relied on supervised learning strategies. Hssayeni et al. [28] recently conducted a comprehensive review of these techniques in both semi-automatic and automatic manners, and binary classification (ICH versus non-ICH) achieved an area-under-the-curve (AUC) of 0.846$\sim$0.975, while more fine-grained ICH subtyping achieved an AUC of 0.93$\sim$0.96. Deep learning-based approaches in ICH detection typically used fully convolutional networks (FCNs) [32] and recurrent neural networks (RNNs) [33], and their accuracy was generally higher for ICH versus non-ICH classification than for ICH subtyping. Following the trend in explainable artificial intelligence (XAI), attention mechanisms have been employed to both boost detection accuracy and visually illustrate classification results.

Saab et al. [64] and Salehinejad et al. [24] utilized ResNet-like architectures for binary ICH detection with attention layers and Grad-CAM techniques, respectively, but they only visualized attention and class activation maps for qualitative assessment of their methods. Furthermore, very limited attempts were also made to apply the attention/class activation in weakly supervised brain lesion and hemorrhage segmentation [34, 35, 65]. Specifically, Wu et al. [34] used refined 3D CAMs to segment stroke lesions from the Ischemic Stroke Lesion Segmentation (ISLES) dataset (multi-spectral MRI), and achieved a 0.3827 mean Dice score. Liu et al. [65] used multi-scale CAMs and a Mixed-UNet model with two decoder branches on top of a VGG-based binary classification CNN. They trained the network based on a private MRI dataset and achieved a 0.56 mean Dice score for ICH segmentation on a small CT dataset. Likewise, Nemcek et al. [35] found the location of ICH as bounding boxes in axial brain CT slices based on the regional extrema of attention maps acquired from a ResNet-like binary classification CNN. In their approach, a mean Dice of 0.58 was reached for the lesion bounding boxes. Unfortunately, to the best of our knowledge, aside from our earlier work [44], self-attention, especially with a Swin transformer, has not yet been explored for weakly supervised ICH segmentation, and we intend to further improve our proposed framework to boost the performance.

4.3 Methods

An overview of our proposed weakly supervised technique for ICH segmentation is depicted in Fig. 4.1, which comprises two major components. First, a Swin transformer was trained through an ICH detection task using categorical labels to classify input images into ICH vs. without ICH. Then, during test time, the segmentation module utilized hierarchical attention maps from the Swin transformer blocks along with their corresponding gradients to predict the hemorrhage segmentation map. Due to the high variability in slice thicknesses among the CT data, we decided to implement our algorithm based on 2D axial slices. The details of the methodology are provided in the following sections.

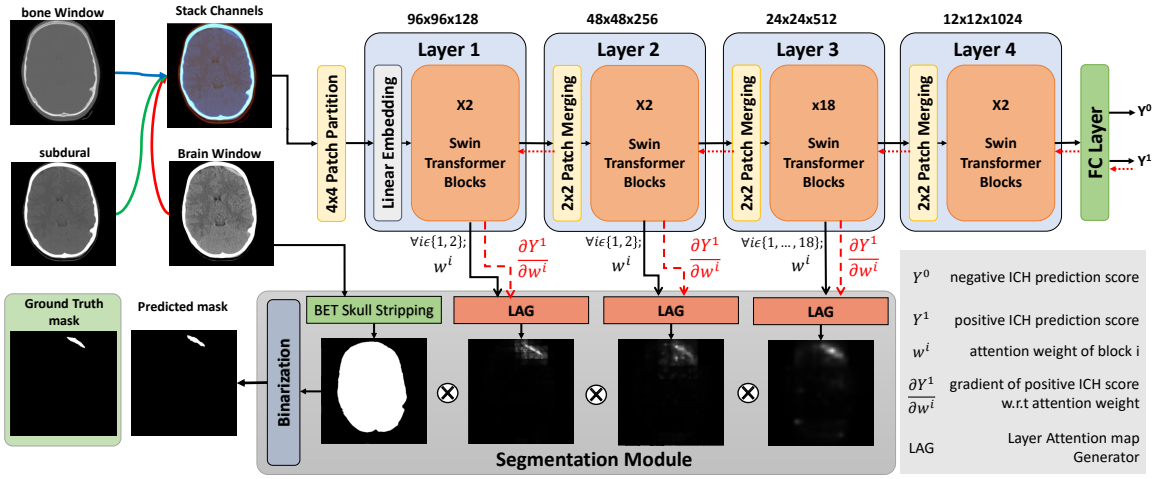

Figure 4.1: Overview of the proposed weakly supervised segmentation method using the hierarchical fusion of gradient-weighted self-attention maps.

4.3.1 ICH detection with a Swin transformer

In our proposed technique, we employed the Swin-Base transformer architecture, which divides an input image into 4×4 patches before passing their embedding through 4 layers/hierarchies to predict the existence of hemorrhages. Unlike the ViT, which computes the multi-head self-attention (MSA) between all image patches, in Swin-Base transformer, self-attention is derived within non-overlapping windows of 12×12 patches, which considerably reduces the computational cost. Here, for simplicity, we will refer to the Swin-Base transformer as "Swin transformer" from this point on. Two main mechanisms help establish the associations between patches across different windows. First, the Patch-Merging module at the beginning of each Swin transformer layer combines and encodes every 2×2 neighboring patches into one. Second, every two consecutive transformer blocks apply window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA) units to input tokens (see Fig. 4.2a). The self-attention per head within each window is computed as:

$$\begin{aligned} \text{Attention}\left(Q_h, K_h, V_h\right) &= w_h \times V_h, \\ w_h &= \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d}} + B_h\right), \end{aligned} \tag{21}$$

where Q, K, and V denote query, key, and value vectors, respectively. w is the window attention weight that we use to derive the attention map, h denotes the head index of multi-head self-attention, d is the dimension of the query or key, and B is the positional embedding matrix. Here, since the dimension of the window is 12×12, the dimension of w is 144×144. Note that w shows the relevance score of key tokens with query tokens. For more information on how the attention weight within each window is computed, we refer the readers to the original Swin transformer paper [15].

In our earlier study [44], we discovered that providing additional information (hemorrhage subtypes) to "ICH vs. without ICH" classification during training can distract the network attention in the Swin transformer. As a result, for our new method with HGI-SAM, we decided to establish the backbone of our algorithm based on simple binary ICH detection. To benefit from the target class gradient, instead of using one output neuron to represent the classification outcome, we framed the final network with a two-class setup (i.e., positive and negative ICH detection). Further information on network training is detailed in Section 4.2.

4.3.2 Hemorrhage Segmentation

In our previous study [44], we have qualitatively demonstrated the superior performance of self-attention maps than the class-activation maps obtained with Grad-CAM in visually explaining the ICH detection process in Swin transformers. Therefore, we continued to take advantage of self-attention maps, with a novel formulation to perform weakly supervised ICH segmentation.

Previous attempts to visualize attention weights in the ViT involved inserting an extra classification token into the image patches and then extracting the attention weight of this token after multiplying the weights of all layers [38, 14]. However, this approach is not feasible for the Swin transformer due to its window division mechanisms for both regular and shifted windows. Additionally, multiplying different attention weights is challenging due to two reasons. First, at different layers/hierarchies, Patch-Merging results in a different feature map resolution and number of tokens. Second, every two successive Swin transformer blocks have attention weights corresponding to regular and shifted image patches that do not match. To address these challenges, we calculated the attention map at each block by averaging over all query tokens with additional operations of the window and shift reversal, and then the interpolated maps at different layers are multiplied.

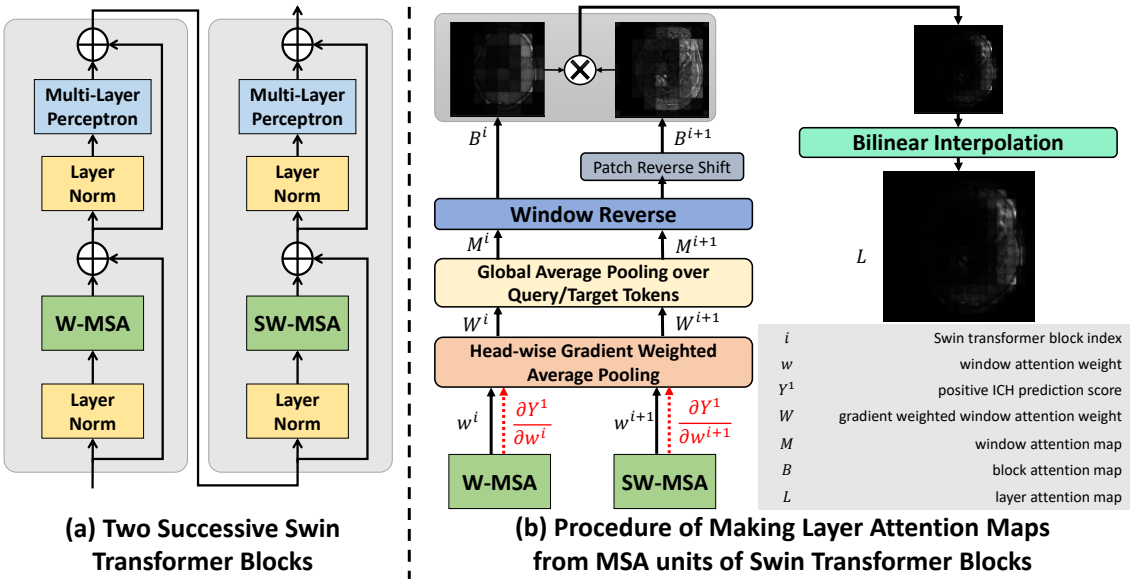

Figure 4.2: Demonstration of head-wise gradient-infused layer attention map generation in the proposed Swin transformer in categorical learning

Layer Attention Map Generation

There has been recent research that leverages model classification scores for attention explainability. For instance, Chefer et al. [38] utilize the Taylor Decomposition principle to assign and propagate a local relevance score through the layers of a ViT model. Similarly, Sun et al. [66] and Barkan et al. [67] employ attention gradient weighting on ViT and BERT models, respectively. However, these approaches primarily focused on the attention weight of the "cls" token, and the latter two methods weighed each token's attention weight through element-wise multiplication. In contrast, our work places emphasis on weighing different heads in multi-head self-attention, and we performed the operation on the more complex Swin Transformer for the first time.

The use of multiple heads in the self-attention mechanism enhances the representational capacity and robustness of the transformer model, as each head can focus on different aspects of the input and learn a unique set of attention weights, thus capturing more complex relationships among the tokens. However, this critical fact was overlooked in most previous attention map generation methods [68], including our own previous work [44]. In the existing literature, naive averaging is often applied to the attention weights of all heads to obtain an overall weight representation. However,

as proved by Voita et al. [69], some heads have more contribution to the output prediction. In this work, we weighed each head by the norm of its gradient regarding the classification score of positive ICH detection, which caused the attention weights of the heads that are more strongly associated with hemorrhage detection to have a heavier influence on the final attention weight representation. This is similar to Grad-CAM, where the target class gradient is used to weigh the associated activation map to enhance its specificity. In a Swin transformer, attention weights are computed within non-overlapping local windows while the W-MSA and SW-MSA units in two successive blocks establish cross-window connections. To encode the full attention information from local windows and cross-window connections, we multiply the attention maps from the original and shifted versions. Thus, we produce one map per every two consecutive blocks. As illustrated in Fig. 4.2, the layer attention map is created as follows:

$$
\begin{aligned}
W^i &= \frac{1}{H}\sum_{h=1}^{H}\left\|\frac{\partial Y^1}{\partial {w^i}_h}\right\| \cdot {w^i}_h, \\
M^i &= \frac{1}{Q}\sum_{k=1}^{Q} W_k^i, \\
L &= \mathrm{BI}\left(\mathrm{WR}\left(W^i\right) \otimes \mathrm{RS}\left(\mathrm{WR}(W^{i+1})\right)\right)
\end{aligned}
\tag{22}
$$

where w^i is the Multi-head window Self-Attention (MSA) weight of block i, H is the number of heads in the MSA unit, M^i is the window attention map of block i which is derived by averaging the window attention weight over its query tokens' dimension, and L is the layer attention map. BI refers to bilinear interpolation, which is utilized to upsample the map to the image size. WR stands for Window Reverse operation, which involves concatenating maps of all windows to create a full image map. Also, RS denotes the reverse shift operation, which is used to reposition the shifted patches of the SW-MSA unit to their original locations in the image. It is good to mention that for Layer 3 in our Swin transformer, which consists of 18 blocks, the final layer map is obtained by averaging the results of 9 maps computed as above.

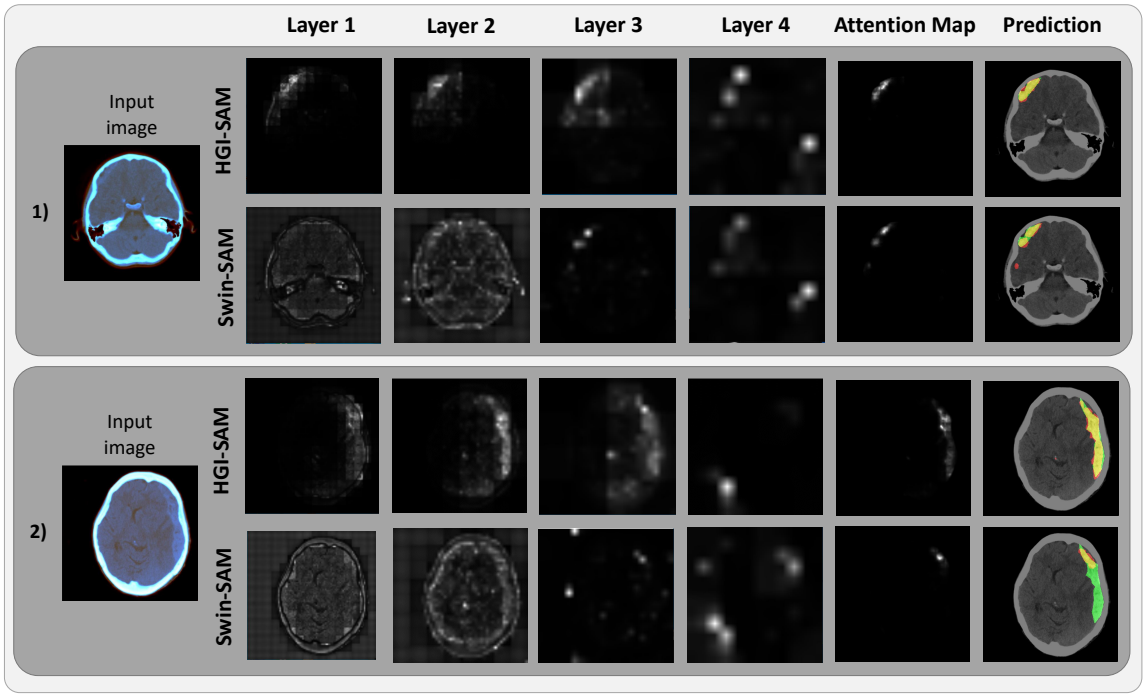

Figure 4.3: A comparison of the proposed head-wise gradient-infused self-attention mapping (HGI-SAM) and the original self-attention maps (Swin-SAM) from the Swin transformer model is shown for two axial CT slices. Along with the maps at different hierarchies, the fused attention maps and the derived binary ICH segmentation (in red) are also shown over the ground truths (in green). Note that the yellow color shows the overlapping area (true positive regions of segmentation results).

Segmentation Module

In the segmentation module, the final ICH segmentation was obtained by thresholding the pixel-wise multiplication result of the attention maps from different hierarchical layers in the Swin transformer. Note that the attention map generated by the last layer in the Swin transformer tends to be much more coarse due to the interpolation of a 12×12 pixel map to the image size, resulting in reduced resolution and potential loss of fine-grained details. Therefore, unlike our previous approach [44], which used the attention of Layer 4 to compensate for its limited ability to capture relevant features in earlier layers, with the new technique using HGI-SAM, we used the attention maps from the first 3 layers to generate the final ICH segmentation [70]. Furthermore, as demonstrated in Fig. 4.1, we employed an additional post-processing step in our approach. This involved multiplying the final fused attention map with a brain binary mask, removing any irrelevant attention weights to ICH segmentation outside the brain region. The skull-stripping procedure was conducted following

the recommended steps outlined by Muschelli et al. [71]. Lastly, the refined attention map was binarized using a simple thresholding method, which was demonstrated to be more robust than K-means or Otsu's method in our previous study [44], resulting in a discrete segmentation mask. To determine the optimal threshold value, we conducted a grid search using the validation data. More specifically, to evaluate a fold in 5-fold cross-validation, we chose a best-performing threshold value from 0 to 1 with a step size of 0.01 that obtained the best segmentation on the remaining folds based on Dice scores.

4.4 Experiments and Evaluation

To investigate the performance of the proposed weakly-supervised ICH segmentation method using our new HGI-SAM technique, in addition to the approaches from our previous publication [44], we also implemented three baseline models, including a fully supervised U-Net, a fully supervised Swin-UNETR, and a similar weakly supervised segmentation method based on binary ICH detection using class activation maps from Grad-CAM. To facilitate the discussion of these methods, we refer the weakly supervised segmentation techniques with Grad-CAM, self-attention maps in multi-label learning (ICH subtyping), self-attention maps in binary ICH detection, and head-wise gradient-infused self-attention maps in binary ICH detection as Swin-Grad-CAM, Swin-SAM Multi-label, Swin-SAM Binary, and Swin-HGI-SAM, respectively. All our networks were trained on a desktop computer with an Intel Core i9 CPU and an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The following sections provide detailed information about the dataset, model training techniques for various segmentation methods, and evaluation metrics.

4.4.1 Dataset

To train and evaluate our models, we used two public datasets, the RSNA ICH CT dataset [42] and the PhysioNet CT dataset [28]. The RSNA dataset contains 752,803 CT slices, with each slice annotated only with ICH subtypes, and the PhysionNet dataset has 2,814 CT slices (75 subjects) with both manual ICH segmentation and ICH subtypes. For all weakly supervised methods, the deep learning models were trained only using the RSNA dataset, which was randomly split into

90% and 10% for training and validation sets. We used the validation set to early stop training when its loss stops decreasing. Their testing was performed only using the PhysioNet dataset. To train and test the fully supervised U-Net and Swin-UNETR networks [47], subject-wise five-fold cross-validation was used on the PhysioNet dataset, where we ensure that no slices from the same subject exist across different folds. Finally, we incorporated the same data splitting to evaluate all techniques. We published our data splitting along with our code at https://github.com/HealthX-Lab/HGI-SAM.

To prepare the data, for each CT slice, brain, subdural and bone windows created using the suggested parameters provided in the relevant data publications [42, 28] were stacked to create a three-channel image, downsampled to 384×384 pixels, and normalized using min-max scaling to the range of [0,1].

4.4.2 Implementation details

ICH segmentation with Swin-SAM Multi-label and Swin-SAM Binary

In our previous work [44], two Swin transformers [36] were trained with categorical learning to provide self-attention maps for ICH segmentation, with one for binary ICH detection and the other for binary ICH detection and full subtyping. When training both models, we used the AdamW optimizer with an initial learning rate of 1e-5 and early stopping with a patience of 3 to avoid overfitting. To address the class imbalance (ICH vs. without ICH) issue in the dataset, we used the focal cross-entropy loss function. Finally, we employed data augmentation techniques, including random left-right flipping, image rotation, and Gaussian noise addition, to improve the capacity and robustness of the trained models. At test time using the PhysioNet data, the binary hemorrhage segmentation was obtained using the same post-processing step as described in Section 3.2.2. More specifically, a five-fold cross-validation approach was used to determine the optimal threshold to generate binary ICH segmentation masks from the fused attention maps. The fold-wise average of threshold values for Swin-SAM Multi-label and Swin-SAM Binary were 0.11 and 0.07, respectively. Additionally, the division of the five folds was made consistent with the training and testing of the supervised U-Net and Swin-UNETR models.

ICH segmentation with HGI-SAM

The Swin transformer for our new weakly supervised ICH segmentation using HGI-SAM was established based on that of the Swin-SAM Binary technique, following our previous insight regarding the benefit of binary classification on self-attention maps [44]. To allow the computation of class-specific gradients for HGI-SAM, instead of one neuron to represent binary ICH detection outcomes, the new model was equipped with two output neurons to represent the ICH positive class and ICH negative class. To take advantage of our existing work, the new model was fine-tuned based on the Swin transformer backbone of Swin-SAM Binary, using the AdamW optimizer with a learning rate of 1e-6 and early stopping. Here, data augmentation with random spatial transformations and Gaussian noise addition was used during training. Furthermore, with the cross-entropy loss function, during training, we adopted data sampling that drew training data samples with probabilities that were inversely proportional to their label frequencies to handle the class imbalance issue in the datasets. Upon completing the training, pixel-wise ICH masks were obtained in the same manner as described in the previous section to allow a consistent comparison for all techniques. The obtained fold-wise average threshold value for this model was 0.06.

4.4.3 Baseline models

ICH segmentation with Grad-CAM

As most existing weakly supervised segmentation techniques relied on Grad-CAM [21], we implemented a baseline technique of this category, where we employed the class activation map on the same Swin transformer that we trained with the binary ICH detection task for Swin-HGI-SAM. Following the suggestion by Jacob et al. [72], we applied the Grad-CAM target layer to the output of the first norm layer in the final block of the Swin transformer. Similar to the proposed self-attention-based method, the activation map was first multiplied by the brain mask and then thresholded to achieve the final hemorrhage segmentation as described in Section 4.2. The obtained fold-wise average threshold value for this model was 0.80.

Fully supervised U-Net

The U-Net is one of the most popular DL models in medical imaging applications. Therefore, we implemented a fully supervised U-Net model with a lighter architecture than that in the PhysioNet ICH data paper [28], which has four hierarchies in the encoding and decoding paths, but less embedding dimension. Each hierarchy consists of two Convolutional layers with ReLU activation function, a Max-Pooling layer in the encoding branch, and Transposed Convolutional layer in the decoding branch. We used the AdamW optimizer with an initial learning rate of 1e-3, the same sampling and augmentation strategy as our weakly supervised models, and a loss function made of Dice coefficient and cross-entropy, in a five-fold cross-validation setup.

Fully supervised Swin-UNETR

Swin-UNETR [47] is one the most popular Swin-based segmentation models that takes advantage of Swin transformer and CNN techniques at the same time. Specifically, it is a U-Net-like architecture, where the encoder is a Swin transformer, the decoder is a CNN, and skip connections pass through convolutional residual blocks. To mitigate overfitting considering the size of the PhysioNet dataset, we adopted a lighter version of its original model that has 4 layers/hierarchies, an initial embedding dimension of 12, and 2, 4, 8, and 16 heads in multi-head self-attention units of Layer 1 to 4. Here, we used the same training parameters and strategies as the U-Net model, which also offers the best outcome for this method, to train the network.

4.4.4 Evaluation metrics

For all the proposed and implemented methods, we evaluated their segmentation performance using Dice coefficient and Intersection over Union (IoU). In addition, to assess the performance of binary ICH detection, we also computed a range of metrics, including accuracy, area under the curve (AUC), precision, F1-score, recall, and specificity for all algorithms. Note that for Swin-SAM Multi-label, where the designated Swin transformer was trained for both binary ICH classification and subtyping, the performance was assessed only based on the binary detection results. For the U-Net and Swin-UNETR models, ICH detection was recorded as whether the network provided a

hemorrhage segmentation for a given image since a similar approach was also used for assessing aneurysm detection in the ADAM MICCAI Challenge [73]. It is worth mentioning that to make the performance of these models in ICH detection more robust, we do not consider tiny foregrounds as positive ICH (< 10 pixels). As the data division for five-fold cross-validations for different techniques was the same, we reported the ICH segmentation and detection accuracy for all folds. We also report the model's overall performance by considering the accuracy of all slices. Lastly, two-sided paired sample t-tests were performed to further confirm the performance of our newly proposed segmentation method based on HGI-SAM against the rest of the comparing group.

4.5 Results

To demonstrate the impact of gradient-weighing for self-attention maps and thus the final hemorrhage segmentation, we illustrate the layer-wise attention maps, along with the combined map and the binary segmentation in Fig. 4.3 for the axial CT slices of two patients. From Fig. 4.3, it is evident that the proposed head-wise gradient-infused self-attention maps (HGI-SAM) provided more attention weights with higher specificity for the hemorrhage regions, especially at the first two layers with higher resolutions. This, in turn, provided final binary segmentations with a better agreement with the ground truths. To showcase the segmentation performance of the proposed method, the results of all mentioned techniques are shown for four different patients in Fig. 4.4. When comparing Swin-Grad-CAM and the self-attention-based results, we can see that while Swin-Grad-CAM could focus on the general region-of-interest correctly, it often provided much larger segmentations than needed. Between Swin-SAM Multi-label and Swin-SAM Binary, as we discovered in the previous study [44], binary classification helped better focus the model attention in the hemorrhage region than the multi-label counterparts, thus offering more accurate segmentation. Finally, in contrast to the rest of the weakly supervised methods, Swin-HGI-SAM gave the most similar results to the fully supervised models, and notably, in Cases 2 and 4, the U-Net missed the small ICH that Swin-HGI-SAM and Swin-UNETR were able to identify.

Following the qualitative demonstration of the segmentation performance, the Dice coefficient and IoU metric for all methods are listed in Table 4.1 for all five folds from the experiments, with

their overall slice-wise mean±SE. While the Swin-UNETR achieved a Dice of 0.455±0.019 and an IoU of 0.355±0.016 in a fully supervised setting, Swin-HGI-SAM was able to offer the second best results, with a 0.444±0.014 Dice. With Swin-Grad-CAM as the worst method, Swin-SAM multi-label and Swin-SAM binary performed worse than the newly proposed technique. In terms of statistical tests for segmentation metrics, Swin-HGI-SAM outperformed all weakly supervised methods ($p = 10^{-37} < 0.05$ compared with Swin-Grad-CAM, $p = 10^{-29} < 0.05$ compared with Swin-SAM Multi-label, $p = 0.0029 < 0.05$ compared with Swin-SAM Binary) while producing similar segmentation accuracy as the fully supervised U-Net ($p = 0.829 > 0.05$) and fully supervised SwinUNETR ($p = 0.6184 > 0.05$)

Finally, in Table 4.2, we listed the full assessment of ICH detection for Swin-SAM multi-label, Swin-SAM binary, Swin-HGI-SAM, U-Net, and Swin-UNETR. Despite the strong performance of fully supervised U-Net and Swin-UNETR models in ICH segmentation, their ICH detection accuracy falls short when compared to weakly supervised models trained with categorical labels. For all Swin transformer models, they offered similar ICH detection performance across all evaluation metrics. By comparing Table 4.1 and Table 4.2 across different data folds, we noticed that the detection results align with segmentation performance, especially for weakly supervised based models. This is expected due to the nature of the proposed weakly supervised segmentation framework.

4.6 Discussion

In recent years, the urgent need to enhance the transparency of deep learning algorithms has encouraged the development of various techniques to visualize network activation/attention maps in vision tasks. Among them, Grad-CAM [21] has gained popularity to reveal the regions of interest in image classification tasks for CNNs, thanks to its simplicity and flexibility. Furthermore, extending its original purpose, it has also been adopted in weakly supervised image segmentation based on categorical and metric learning to generate pixel-level semantic labels [60], including applications for stroke lesion segmentation [34, 35]. Compared with Grad-CAM and its variants, the more recent attention mechanisms, especially self-attention from transformer models, can identify more

Table 4.1: Assessment of ICH segmentation performance for Swin-Grad-CAM, Swin-SAM Multi-label, Swin-SAM Binary, Swin-HGI-SAM, U-Net, and Swin-UNETR algorithms, using Dice coefficient and Intersection over Union (IoU). All results are reported as mean±SE. Note the overall metrics are reported based on all cases across the folds.

Fold	Dice Coefficient					
	Swin-Grad-CAM	**Swin-SAM Multi-label**	**Swin-SAM Binary**	**Swin-HGI-SAM**	**Fully supervised U-Net**	**Fully supervised Swin-UNETR**
1	0.174 ± 0.025	0.223 ± 0.029	0.337 ± 0.031	0.354 ± 0.040	0.302 ± 0.045	0.281 ± 0.044
2	0.219 ± 0.024	0.319 ± 0.021	0.433 ± 0.025	0.505 ± 0.026	0.336 ± 0.036	0.434 ± 0.044
3	0.189 ± 0.032	0.276 ± 0.032	0.347 ± 0.033	0.414 ± 0.034	0.442 ± 0.047	0.399 ± 0.045
4	0.268 ± 0.022	0.322 ± 0.025	0.400 ± 0.025	0.451 ± 0.029	0.555 ± 0.036	0.571 ± 0.034
5	0.307 ± 0.026	0.338 ± 0.029	0.407 ± 0.022	0.481 ± 0.030	0.491 ± 0.040	0.522 ± 0.037
Overall slice-wise	**0.237 ± 0.012**	**0.299 ± 0.012**	**0.387 ± 0.012**	**0.444 ± 0.014**	**0.438 ± 0.019**	**0.455 ± 0.019**

Fold	Intersection over Union					
	Swin-Grad-CAM	**Swin-SAM Multi-label**	**Swin-SAM Binary**	**Swin-HGI-SAM**	**Fully supervised U-Net**	**Fully supervised Swin-UNETR**
1	0.107 ± 0.017	0.143 ± 0.020	0.226 ± 0.024	0.255 ± 0.031	0.228 ± 0.036	0.210 ± 0.035
2	0.136 ± 0.017	0.200 ± 0.015	0.295 ± 0.020	0.360 ± 0.022	0.238 ± 0.029	0.338 ± 0.038
3	0.127 ± 0.023	0.185 ± 0.024	0.240 ± 0.026	0.294 ± 0.027	0.352 ± 0.041	0.308 ± 0.038
4	0.171 ± 0.015	0.214 ± 0.018	0.275 ± 0.019	0.328 ± 0.024	0.453 ± 0.033	0.458 ± 0.030
5	0.199 ± 0.019	0.226 ± 0.022	0.270 ± 0.018	0.347 ± 0.026	0.381 ± 0.035	0.403 ± 0.033
Overall slice-wise	**0.151 ± 0.008**	**0.196 ± 0.009**	**0.263 ± 0.010**	**0.319 ± 0.012**	**0.343 ± 0.016**	**0.355 ± 0.016**

Table 4.2: Assessment of ICH detection performance for Swin-SAM Multi-label, Swin-SAM Binary, Swin-HGI-SAM, U-Net, and Swin-UNETR algorithms, using accuracy, AUC, precision, F1-score, recall, and specificity. Note the overall metrics are reported based on all cases across the folds.

Fold	Accuracy					AUC				
	Swin-SAM Multi-label	**Swin-SAM Binary**	**Swin-HGI-SAM**	**U-Net**	**Swin-UNETR**	**Swin-SAM Multi-label**	**Swin-SAM Binary**	**Swin-HGI-SAM**	**U-Net**	**Swin-UNETR**
1	0.948	0.953	0.946	0.572	0.654	0.821	0.874	0.904	0.731	0.785
2	0.958	0.964	0.958	0.751	0.816	0.851	0.891	0.902	0.830	0.845
3	0.934	0.928	0.928	0.674	0.589	0.712	0.701	0.731	0.765	0.732
4	0.967	0.965	0.953	0.692	0.685	0.932	0.950	0.948	0.805	0.810
5	0.959	0.954	0.938	0.515	0.660	0.935	0.939	0.937	0.721	0.788
Overall slice-wise	**0.953**	**0.953**	**0.945**	**0.639**	**0.679**	**0.858**	**0.879**	**0.891**	**0.770**	**0.793**

Fold	Precision					F1-score				
	Swin-SAM Multi-label	**Swin-SAM Binary**	**Swin-HGI-SAM**	**U-Net**	**Swin-UNETR**	**Swin-SAM Multi-label**	**Swin-SAM Binary**	**Swin-HGI-SAM**	**U-Net**	**Swin-UNETR**
1	0.735	0.724	0.657	0.166	0.201	0.699	0.750	0.742	0.282	0.331
2	0.894	0.870	0.803	0.304	0.369	0.792	0.832	0.817	0.458	0.520
3	0.862	0.800	0.737	0.226	0.191	0.575	0.545	0.583	0.359	0.315
4	0.893	0.849	0.784	0.322	0.319	0.888	0.888	0.856	0.482	0.483
5	0.767	0.731	0.652	0.183	0.238	0.830	0.814	0.768	0.308	0.381
Overall slice-wise	**0.830**	**0.796**	**0.725**	**0.231**	**0.253**	**0.780**	**0.789**	**0.770**	**0.370**	**0.398**

Fold	Recall (Sensitivity)					Specificity				
	Swin-SAM Multi-label	**Swin-SAM Binary**	**Swin-HGI-SAM**	**U-Net**	**Swin-UNETR**	**Swin-SAM Multi-label**	**Swin-SAM Binary**	**Swin-HGI-SAM**	**U-Net**	**Swin-UNETR**
1	0.667	0.778	0.852	0.926	0.944	0.976	0.970	0.956	0.537	0.625
2	0.712	0.797	0.831	0.932	0.881	0.989	0.985	0.974	0.728	0.808
3	0.431	0.414	0.483	0.879	0.914	0.992	0.988	0.980	0.651	0.551
4	0.882	0.929	0.941	0.965	0.988	0.982	0.971	0.955	0.645	0.632
5	0.903	0.919	0.935	0.984	0.952	0.966	0.958	0.938	0.457	0.624
Overall slice-wise	**0.736**	**0.783**	**0.821**	**0.940**	**0.940**	**0.981**	**0.974**	**0.960**	**0.600**	**0.645**

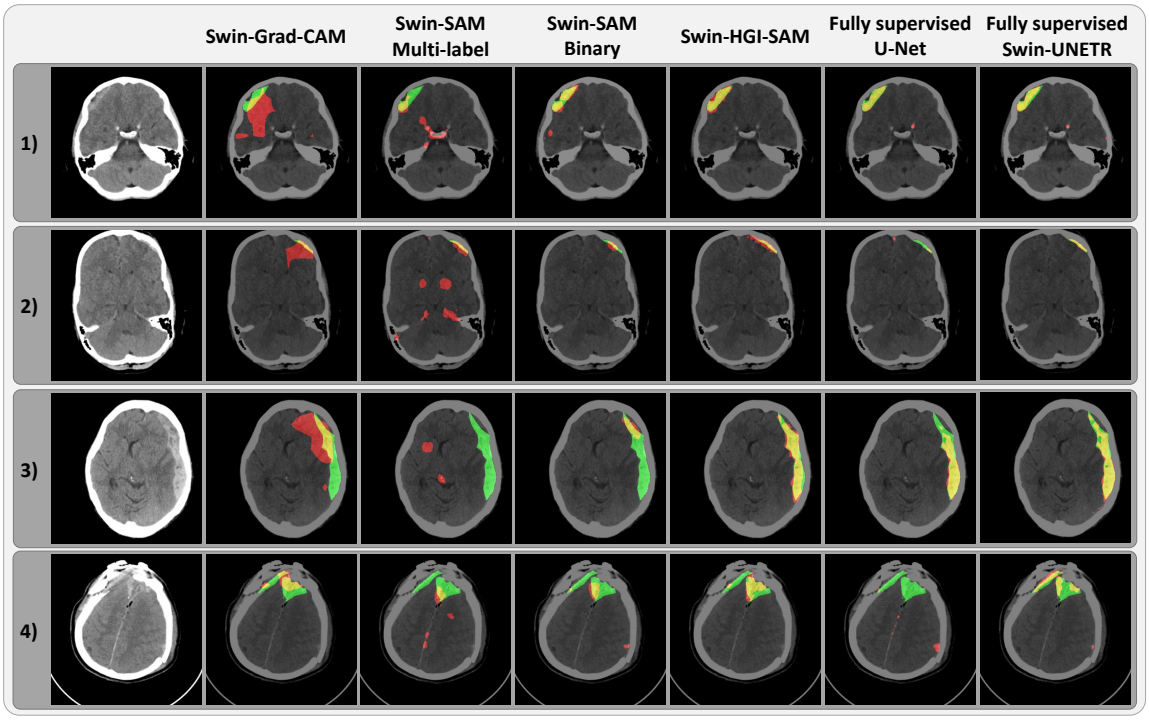

Figure 4.4: Qualitative comparison of segmentation performance for the proposed weakly supervised ICH segmentation methods (Swin-Grad-CAM, Swin-SAM Multi-label, Swin-SAM Binary, and Swin-HGI-SAM), fully supervised U-Net, and fully supervised Swin-UNETR for four different cases. Here, red=automatic segmentation, green=ground truths, and yellow=true positives.

discriminative, task-related regions and features while improving the performance of the DL models [74, 14]. This was confirmed in this study when comparing the segmentation performance of the proposed weakly supervised ICH segmentation approaches with Grad-CAM and self-attention maps. As for the self-attention mechanism, different learning strategies may influence the positioning and tightness of network attention with respect to the target objects, and thus the downstream segmentation outcomes in the proposed framework. As an ablation study of our previous work [44], we examined the impact of binary (presence of hemorrhage or not) versus multi-label classifications (ICH subtypes and with/without ICH) on self-attention maps from Swin transformers. Using segmentation accuracy as a metric, we found that binary classification helped the network better focus on the hemorrhage regions while both strategies offer similar performance for ICH detection. In the new segmentation method with HGI-SAM, we followed our earlier insights to build our algorithm.

Inspired by the popular Grad-CAM technique [21], we incorporated head-wise gradient-weighing

for self-attention maps to boost the presentation of the weights relevant to specific class activation for the first time. Compared with other attention mapping techniques [38, 66] that relied on the ViT, we were also the first to implement it on the more complex Swin transformer that was intended to improve upon the ViT. The enhanced visualization of the attention maps and ICH segmentation accuracy are evident in Fig. 4.3 and Fig. 4.4, respectively. Among the obtained attention maps at different hierarchies, those from the earlier layers contained the relevant attention weights at higher resolutions, and thus were more helpful to delineate the regions of interest (i.e., hemorrhages) through the ICH classification task. In our experiment, the head-wise gradient-infused self-attention map (HGI-SAM) from Layer 4 possesses relatively less discriminative power primarily due to much lower resolution and the adverse cascading effect of transformer architecture [70]. Therefore, we chose to fuse those from the first three layers for our proposed method. In fact, with a few cases, we found that the fused attention map from the first 3 layers offered better segmentation accuracy than using those from all 4 layers. In our original study that relied on self-attention maps alone, fusing all 4 layers was more beneficial. To obtain discrete ICH segmentation from the HGI-SAM, we applied additional post-processing steps. One major procedure that was different from our original article was multiplying the brain mask before ICH mask binarization. When closely inspecting the attention maps from ICH classification, we noticed that skull fractures were also identified in addition to the hemorrhage. This is likely because, for many ICH patients, the condition may result in accidental falls that cause additional injury, such as skull or spine fractures. This phenomenon perfectly showcased the power of attention visualization in explaining the decision-making process in DL models. By constraining the post-processing in the brain region, we intended to exclude the attention weights regarding skull fractures and were able to further improve the segmentation accuracy. Finally, different from our previous approach [44], where the denoised brain window was multiplied to the attention map, our new method directly performed thresholding on the gradient-weighted map to avoid potential intensity inconsistency within the hemorrhage and multi-center imaging protocols. This also allowed us to directly probe the quality of activation/attention maps with respect to their specificity in focusing on ICH.

To provide baselines for our weakly supervised segmentation framework, we have trained a

U-Net and a Swin-UNETR with full supervision using the PhysioNet data to perform ICH segmentation. By using data sampling to tackle class imbalance in training, our U-Net model has achieved an improved mean Dice score of 0.438 over that of 0.315 reported for the U-Net in the original data paper [28]. In comparison, our proposed method has achieved similar results to our baseline supervised U-Net and Swin-UNETR ($p > 0.05$) with the mean Dice scores slightly lower than the Swin-UNETR and higher than the U-Net, showcasing the feasibility and excellent potential of weakly supervised segmentation with much more accessible categorical labels. In terms of computational cost, U-Net was the most efficient model, taking only around 10ms/sample, likely due to its simple convolutional layer architecture. On the other hand, Swin-UNETR took around 15ms/sample, Swin-SAM models took around 30ms/sample, and Swin-HGI-SAM and Swin-Grad-CAM took approximately 60ms and 90ms per sample, respectively. The longer inference time is because the latter two required backward operations for gradient computation, which is a key step for the proposed framework. However, all these models are still relatively fast and suitable for clinical setups, offering practical benefits.

While there is still room for improvement in our future work, ICH segmentation from clinical scans remains a challenging task at the moment. In our proposed framework, extracting meaningful pixel-wise attention maps is crucial. We admit that the exploration of self-attention in this study may be data-, application- and model-specific while the baseline supervised models have been tested in various applications. By using categorical learning to obtain attention and saliency maps for segmentation, depending on the data and application, it is possible that the local regions that the network focuses on for image classification may not fully overlap with the segmentation ground truths. In our application, the derived self-attention maps focused on both ICH lesions and skull fractures in some cases, and we used skull stripping to tackle this. In the future, we will continue to investigate the characteristics of self-attention in different learning strategies, extended applications, and other Transformer models. These would be greatly beneficial to improve weakly supervised medical image segmentation based on categorical labels. Incorporating inter-slice or 3D spatial information may be beneficial to the designated tasks, especially for 2D slices that contain a few pixels of ICH, but the high variability of CT slice thickness in the public datasets has posed

challenges in the 3D approach. Recent developments in resolution-agnostic brain image segmentation [75] and image super-resolution [76] through generative DL models have allowed high-quality interpretation of clinical scans with diverse imaging protocols (e.g., different image resolutions). We will seek to adapt these frameworks for CT images in the task of ICH detection and segmentation in future work.

4.7 Conclusion

To mitigate the requirement of expensive training data for intracranial hemorrhage segmentation, we have proposed a weakly supervised framework by using a novel hierarchical combination of head-wise gradient-infused self-attention maps from a Swin transformer through categorical learning. By using two public CT databases, we further demonstrated the benefits of head-wise gradient-weighing of derived attention maps to further boost ICH segmentation performance for the first time. In the future, we will further explore the proposed HGI-SAM technique and the application of the proposed weakly supervised segmentation framework in extended applications and other Transformer models.

Chapter 5

Intracranial Aneurysm Segmentation Using A Novel Focal Modulation UNet and Conditional Random Fields

A version of this chapter is uploaded to Arxiv as a preprint, and submitted to the 15th International Conference on Information Processing in Computer-Assisted Interventions.

- **A. Rasoulian**, S. Salari, Y. Xiao, "Weakly supervised segmentation of intracranial aneurysms using a 3D focal modulation UNet," ArXiv:2308.03001, 2023.

5.1 Introduction

An intracranial aneurysm is an abnormal focal bulging or ballooning of the vasculature in the brain due to weakened blood vessel walls. Although often asymptomatic, large aneurysms can cause headaches, visual disturbance, neurological deficits, and other issues [1]. Furthermore, a ruptured aneurysm can lead to life-threatening subarachnoid hemorrhage, a severe stroke subtype. Typical treatments for unruptured intracranial aneurysms (UIAs) include endovascular coiling and surgical clipping. When screening for UIAs and determining the treatment plans, risk assessment of UIA growth and rupture is crucial. Thanks to modern medical imaging techniques, such as computed

tomography (CT) and magnetic resonance imaging (MRI), the ability to accurately and reliably identify and measure aneurysms can ensure therapeutic outcomes. Compared with CT angiography (CTA), time-of-flight magnetic resonance angiography (TOF-MRA) does not expose the patients to radiation or adverse reactions toward contrast agents. It thus is better suited for routine follow-up imaging.

Manual identification and measurements of UIAs can be difficult and time-consuming, especially for those that are of small sizes (<5mm), and it is estimated that 10% of all UIAs are missed during screening [77]. Therefore, automatic algorithms that allow detection and especially 3D segmentation of the aneurysms from medical scans can greatly facilitate the clinical workflow and enable more fine-grained risk analysis based on UIA shapes. To date, many techniques [78, 79] have been proposed for UIA detection, but only a few reported performance for 3D UIA segmentation from CTA/MRA. Notably, the ADAM Challenge [78] hosted in conjunction with MICCAI 2020 attracted nine DL-based UIA segmentation algorithms. Among these, various approaches of convolutional neural networks (CNNs), especially variants of the UNet have been proposed, with the winning algorithm achieving an average Dice score of 0.64 and a mean 95% Hausdorff distance (95-HD) of 2.62mm for correctly detected UIAs. However, most existing techniques were developed based on private clinical data with refined manual annotations, and many focused on CTA. Further investigation is still needed for this challenging task, especially for MRA with weaker contrast. The lack of large well-annotated databases, which are costly to acquire and demand high domain expertise, poses significant challenges in developing deep learning methods in medical imaging applications, particularly for anatomical segmentation. To mitigate this, weakly supervised segmentation [18, 80, 44, 81] leverages coarse annotations that are easier to obtain (e.g., rough segmentation and categorical image labels) to derive refined segmentation. In this paper, we employ coarse segmentation ground truths of UIAs from TOF-MRA to detect and segment the pathology by using a novel 3D focal modulation UNet, called FocalSegNet, and post-processing with a fully connected conditional random field (CRF) model. In summary, our work has three major novel contributions. **First**, inspired by the recent focal modulation technique [16] and the UNet architecture [82], we proposed a novel 3D UNet with focal modulation; **Second**, we thoroughly compared the performance of self-attention [13] and focal modulation [16] in weak segmentation of UIAs; **Lastly**, we revealed the

key factors that contribute to UIA segmentation for the proposed method through various ablation studies.

5.2 Methodology

5.2.1 Network architecture of FocalSegNet

While 3D UNet was a common choice for UIA detection and segmentation [78], the newer Vision Transformer (ViT) and its variants (e.g., Swin Transformer) that leverage self-attention mechanisms [13, 15] have become increasingly popular in medical imaging applications, providing superior performance. More recently, to better encode contextual information with a lighter model, Yang et al. [16] proposed focal modulation for DL in vision tasks, and demonstrated that it performed better than the Swin Transformer. In focal modulation networks, self-attention is replaced by a stack of depth-wise convolutional layers that focally encode visual contexts and selectively gather them into a modulator using a gated aggregation. Next, the modulator is injected into the query and passed to the next block. Both self-attention and focal modulation techniques involve linearly determining the key, query, and value based on input tokens. Nevertheless, the primary distinction lies in their operational sequence: in self-attention, a query-key interaction occurs before aggregation with the value, whereas in focal modulation, the value is first aggregated with contexts around each key, and subsequently, the result is adaptively modulated with the query. Despite excellent performance in classification and segmentation of natural images, focal modulation's performance in 3D medical images is yet to be verified. Recently, Naderi et al. [83] proposed a 2D UNet-like architecture with focal modulation blocks as the encoder and decoder for abdominal CT segmentation. It offered a higher mean Dice score (not confirmed by statistical tests) than the Swin-Unet [46], another UNet-like model with Swin Transformers as the encoder and decoder. Different from [83], we followed the approach of the recent 3D Swin-UNETR [47], a popular Swin-CNN hybrid model to build a new architecture by replacing the Swin Transformer with 3D focal modulation for the encoder. At the encoder branch, each layer's output, consisting of image tokens and their embeddings, is reshaped to a volumetric patch and passed through a residual block to work as skip connections. Starting from the bottleneck, each feature map is expanded using a transposed convolutional block and then

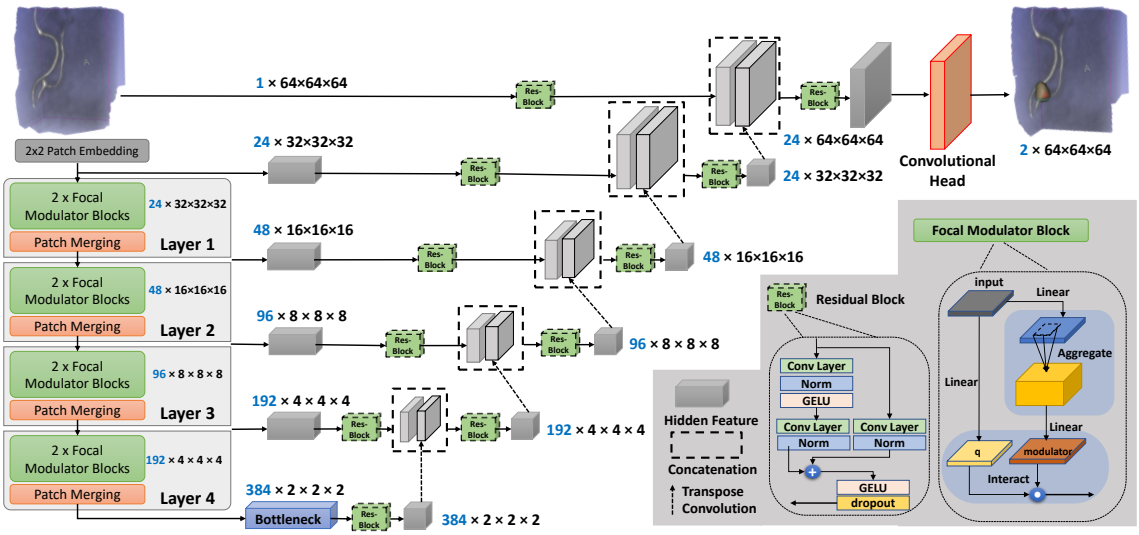

Figure 5.1: Network architecture of the proposed FocalSegNet

concatenated with the corresponding skip connection before passing through another residual block to the next layer. The detailed architecture is depicted in Fig. 5.1.

5.2.2 Post-processing with fully connected CRF

As the proposed deep learning (DL) models (i.e., FocalSegNet) were trained on rough segmentations that overestimate the true aneurysm shapes, we used a fully connected CRF [84] to further refine our model's initial prediction ($P(x_i)$). If we define X_i as a random variable representing the assigned label to pixel i (foreground/background), and I_i as a global observation characterizing the pixel information in the input image, the CRF is modeled as the pair of (I, X) which follows a Gibbs distribution of the form $P(X = x|I) = \frac{1}{Z(I)} exp(-E(x|I))$ [85]. Here, $Z(I)$ is a normalization factor, and $E(x|I)$ is the energy function of assigning labels x to the image. Through minimizing this energy function, the new labels will have a greater likelihood of being assigned to the image. Inspired by [17], we define the energy function as:

$$E(x|I) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{23}$$

$$\psi_u(x_i) = -logP(x_i)$$

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \left[\omega_1 exp(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}) + \omega_2 exp(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}) \right]$$

$$\mu(x_i, x_j) = 1 \text{ if } x_i \neq x_j \text{ and } 0 \text{ otherwise}$$

Here, $\psi_u(x_i)$ is the unary term measuring the cost of assigning label x_i to pixel i, and $\psi_p(x_i, x_j)$ is the pairwise term that shows the cost of assigning labels x_i and x_j to pixels i and j simultaneously. The unary term is derived from our FocalSegNet output logits, and the pairwise term is computed by applying a bilateral kernel and a Gaussian kernel on the original MRA patch image. The first kernel forces the pixels with similar positions (p) and intensities (I) to have similar labels, and the second enforces the smoothness of prediction by only considering the spatial proximity. ω_1 and ω_2, σ_α, σ_β, and σ_γ are hyper-parameters that control the weight and scale of these kernels and are chosen empirically. We used connected-component analysis [86] to identify different clusters when the algorithms predict multiple UIAs, and each connected component was treated as an individual detected aneurysm. To further remove noise, any cluster that had lower than 10 voxels was discarded.

5.3 Experiments and Evaluation

5.3.1 Dataset and pre-processing

Unfortunately, we were not given access to the ADAM challenge dataset [78]. Instead, we relied on a publicly accessible TOF-MRA dataset of UIAs described in [87] to develop and validate our algorithms. It contains 284 subjects (170 females, 127 healthy controls/157 patients with 198 aneurysms). For 246 subjects, coarse spheres that enclose a whole aneurysm were manually labeled, while 38 subjects have voxel-wise segmentation. All scans were pre-processed with skull-stripping and bias field correction [88], and finally resampled to a median resolution of $0.39 \times 0.39 \times 0.55$ mm^3. To allow efficient computation, we performed segmentation based on 3D image patches instead of the entire brain volume. Furthermore, we followed the "anatomically informed" approach in [87] to extract image patches of $64 \times 64 \times 64$ voxels guided by a probabilistic cerebrovascular atlas with anatomical landmarks on the most probable locations of having an aneurysm, resulting in

$\sim$50 patches per subject. In summary, several positive patches with different offsets around every aneurysm were extracted, and for negative patches, a variety of vessel-like, landmark-centered, and random patches were obtained. For more details and the script for image patch extraction, we refer the readers to the original data paper [87]. It is worth mentioning that we used the same patch-extracting strategy for train and test sets. Finally, for DL model training and inferencing, all image patches were normalized with z-transformation. Note that the cases with coarse labels were divided subject-wise for training (95%) and validation (5%), and those with refined segmentation masks were saved for model testing only.

5.3.2 Implementation details and baseline models

In our proposed FocalSegNet, we utilized a 3D version of the original Focal modulation network [16] as the encoder. Each layer of the encoder consists of two Focal Modulator Blocks, comprising hierarchical depth-wise convolutional layers with kernel sizes of 3, 5, and 7, alongside GELU activation functions. For the baseline models to compare against the FocalSegNet, we also implemented a 3D Residual-UNet that unlike the simple UNet used in the original data paper [87], takes advantage of residual connections for better gradient and information flow, and a Swin-UNETR with default parameters [47] to compare two similar mechanisms, self-attention and focal modulation. The only difference between Swin-UNETR and FocalSegNet is their encoders, which makes our comparison valid. All networks had four layers/hierarchies, equal embedding dimensions, and were trained with a batch size of 12 and an AdamW optimizer (weight decay for L2 regularization=1e-6 and initial learning rate=1e-3). We used a step learning scheduler that reduces the learning rate by 2% every 100 iterations. Furthermore, we used sampling, data augmentation, and a combination of different loss functions to tackle the class imbalance problem. Early stopping is also used to avoid model overfitting.

Sampling and data augmentation:

The ratio of patches without and with UIAs is 9:1, which can cause the model to be biased toward the negative class. To tackle this, each patch in the training set is assigned a probability distribution inversely proportional to its label's frequency of occurrence. Then, each batch is randomly sampled

from this distribution with replacement, ensuring the number of positive and negative patches in a batch is almost the same. Although some samples may not be seen in one epoch, there is a high likelihood that the network will see all of them after several epochs. We also use online image augmentation techniques, including random flipping, rotation of up to 15 degrees along x-, y-, and z-directions, and the addition of random Gaussian noise to mitigate overfitting.

Loss function

Since the small sizes of UIAs can cause a large class imbalance, we composed our total loss function with likelihood-based cross-entropy loss (CE), regional-based generalized Dice loss (GD) [89], and distance-based boundary loss (Boundary) [90] as:

$$Loss = \alpha CE + \beta GD + \gamma Boundary \tag{24}$$

$$CE = -\sum_{c} g_c \cdot log(p_c)$$

$$GD = 1 - 2 \times \sum_{c} \omega_c \frac{p_c \odot g_c}{p_c + g_c}$$

$$Boundary = p^1 \cdot \phi(g^1)$$

where g and p are ground-truth and prediction logits of a pixel, respectively. The ω is a weight assigned to each class inversely proportional to its frequency, and $\phi(g^1)$ is the distance map of the foreground class, and its value at each pixel equals the Euclidean distance between that point and the closest background point. $\alpha = \beta = 2\gamma = 0.40$ were determined empirically. These individual loss functions help improve the detection and segmentation of UIAs for the proposed and baseline models.

5.3.3 Evaluation metrics

For the outcomes of the proposed algorithm and its counterparts, as well as the associated ablation studies, we included five different metrics to evaluate their performance in detecting and segmenting the aneurysm based on the test dataset on a per-image-patch basis. Specifically, for UIA detection

quality, we measured the sensitivity and false positive (FP) rate. Note that within the same image patch, it is feasible to contain multiple UIAs. Thus, we separated distinct connected components in the prediction map and ground truth, and evaluated them per aneurysm. Here, a correct aneurysm detection is defined as when the centroid of the predicted segmentation component lies within the boundary of an aneurysm and is considered a false positive otherwise. Next, for all the correctly identified aneurysms, we further evaluated the segmentation accuracy using the Dice coefficient, Intersection over Union (IoU), and 95-$\%$ Hausdorff distance (95-HD) measured in voxels to verify the quality of region and boundary overlaps. Because an aneurysm may be detected by one model and missed by the other, we performed an unpaired two-sided t-test to compare the performance between different experimental setups. A p-value < 0.05 was used to indicate a significant difference.

5.4 Results

5.4.1 UIA detection and segmentation

The UIA detection and segmentation performance for all three DL models (i.e., Residual-UNet, Swin-UNETR, and FocalSegNet) with CRF finetuning is listed in Table 5.1, with segmentation results for two subjects illustrated in Fig. 5.2. In terms of aneurysm detection, for FP rate, our FocalSegNet well outperformed the other two networks ($p < 0.05$), while for sensitivity, all methods perform similarly ($p > 0.05$). This implies that FocalSegNet is better at distinguishing true/false UIAs, which could be hard even for human raters at times. Although FocalSegNet has the best mean segmentation metrics, the difference with Swin-UNETR was not significant ($p > 0.05$). However, both outperform the Residual-UNet ($p < 0.05$).

5.4.2 Ablation studies

Besides comparing the performance of the full setup, where fully-connected CRF was used to refine the initial results of the DL models, including FocalSegNet, Swin-UNETR, and Residual-UNet, we also conducted ablation studies to test the impact of CRF post-processing, as well as each component of the full loss function for the proposed FocalSegNet. We evaluated the metrics for UIA detection and segmentation for all the experiments to gain the required insights.

Table 5.1: UIA detection and segmentation performance (mean±std) of different deep learning models with and without CRF post-processing.

	FP rate	**Sensitivity**	**Dice**	**IoU**	**95-HD** (voxels)
Residual-UNet	0.322 ± 0.537	0.793 ± 0.405	0.587 ± 0.144	0.430 ± 0.140	2.870 ± 1.262
Residual-Unet + CRF	0.277 ± 0.500	0.778 ± 0.415	0.668 ± 0.129	0.515 ± 0.137	2.315 ± 1.125
Swin-UNETR	0.476 ± 0.596	0.866 ± 0.340	0.625 ± 0.137	0.468 ± 0.137	2.495 ± 0.957
Swin-UNETR + CRF	0.403 ± 0.560	0.841 ± 0.366	0.668 ± 0.144	0.518 ± 0.149	2.214 ± 1.032
FocalSegNet	0.231 ± 0.488	0.827 ± 0.379	0.638 ± 0.130	0.481 ± 0.132	2.504 ± 1.140
FocalSegNet + CRF	0.212 ± 0.464	0.801 ± 0.399	0.677 ± 0.141	0.527 ± 0.144	2.148 ± 1.082

From Table 5.1, we observe that CRF filtering has a positive impact in boosting the accuracy of the initial segmentation of all models ($p < 0.05$) while FocalSegNet offers the best performance without CRF ($p < 0.05$). Table 5.2 demonstrates the impact of design choices in terms of the loss function on FocalSegNet. The network trained solely with the Cross-Entropy loss function performs poorly due to the aneurysm size being negligible compared to background voxels, leading to a bias toward negative predictions. Incorporating Generalized Dice Loss enhances performance, but the false positive (FP) rate remains high. To address this, boundary loss is added, enforcing closer proximity between predictions and ground truths. Ultimately, post-processing with a CRF further enhances segmentation performance.

Table 5.2: Influence of loss functions and CRF post-processing on the proposed FocalSegNet in UIA detection and segmentation results (mean±std)

	FP rate	**Sensitivity**	**Dice**	**IoU**	**95-HD** (voxels)
Cross-Entropy loss	0.002 ± 0.048	0.006 ± 0.075	0.018 ± 0.007	0.009 ± 0.004	4.215 ± 0.121
+ Generalized Dice loss	0.527 ± 0.729	0.864 ± 0.343	0.604 ± 0.143	0.447 ± 0.144	2.854 ± 1.291
+ Boundary loss	0.231 ± 0.488	0.827 ± 0.379	0.638 ± 0.130	0.481 ± 0.132	2.504 ± 1.140
+ CRF	0.212 ± 0.464	0.801 ± 0.399	0.678 ± 0.141	0.527 ± 0.144	2.148 ± 1.082

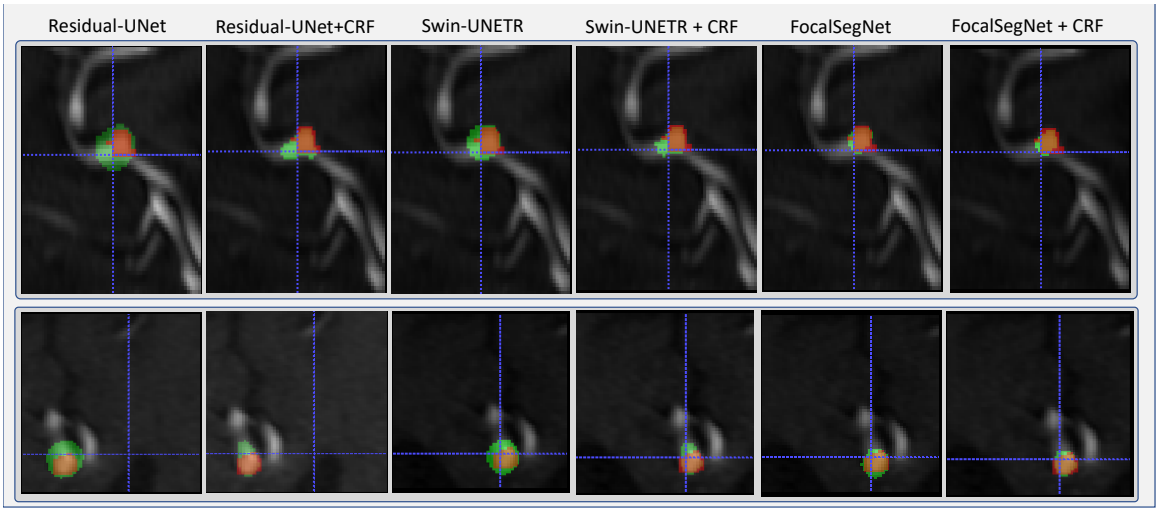

Figure 5.2: Comparison of segmentation results of different techniques for two different patients (one patient per row).Red label=ground truths and green label=automatic segmentation.

5.5 Discussion

Cerebral aneurysm segmentation is still challenging due to its high similarity in adjacent vessel structures and typically small sizes in contrast to the full brain volume. As previous reports [87, 78] primarily focus on binary detection and/or localization of cerebral aneurysms, the relevant reports on the assessment of segmentation algorithms are very limited. Many earlier primary works rely on contrast-enhanced CT angiographies, which have sharper contrast for vasculatures than TOF-MRA, and in-house collected data with refined annotations. Compared with the best segmentation results from the ADAM Challenge (for detected true UIAs) with a 0.64 mean Dice score and 95-HD of 2.62 mm [78], our proposed FocalSegNet, together with CRF post-processing has achieved a Dice score of 0.677 ± 0.141 and a 95-HD of 2.15 voxel ($\sim$0.95mm), only by leveraging rough voxel-wise annotations of the aneurysms. Regarding UIA detection/localization, our proposed method obtained a sensitivity of 0.801 and an FR rate of 0.212, which are also excellent results. Note that since our UIA detection performance was assessed on a per image-patch basis, the resulting metrics may not be directly comparable to some of those previous reports that were done subject-wise.

In almost all previous reports on the relevant topic, the UNet and its variants have been widely adopted. Here, we employed the 3D Residual-UNet as a baseline model to properly assess the segmentation quality against the FocalSegNet and Swin-UNETR, whose general architectures were

also inspired by the UNet, but with the encoding branch modified with Transformer or focal modulation blocks, respectively. This comparison also allows us to probe the characteristics and performance of these two new DL techniques that model long-range contextual information through different mechanisms. For our application in a 3D weakly supervised segmentation, we confirm the claim of [16] for the benefit of focal modulation, with Swin-UNETR ranking the second. This is likely because aneurysms, especially the small ones, are often attached to the main arteries near the branching points, and thus contextual knowledge of anatomy will be beneficial. This is in addition to the anatomical prior-based approach that we adopted from Di Noto et al. [87] in training data sampling.

In terms of GPU memory usage during training, FocalSegNet, Swin-UNETR, and Residual-UNet took 11.2GB, 17.4GB, and 9.7GB respectively, based on the batch size of 12. For all networks, the inference time per patch is around 70ms based on a desktop computer with Intel Corei9 CPU, Nvidia GeForce RTX 3090 GPU, and 24GB RAM. This offers a glance of the efficiency of the proposed network. Compared to [16, 83], we further extended the 2D focal modulation to 3D in a new UNet-like architecture for segmentation tasks for the first time. In future works, we will further examine the proposed FocalSegNet for other supervised and weakly supervised segmentation tasks in other anatomical structures.

For all the DL models, we used a combination of different loss functions, including cross-entropy loss, generalized Dice loss, and boundary loss. While generalized Dice loss and cross-entropy loss have been popular for many medical image segmentation tasks, boundary loss [90] is helpful in segmenting structures with high-class imbalance, such as the case of aneurysms segmentation. With the intention to reduce the reliance on carefully annotated manual segmentation, many have attempted different strategies in weakly or semi-supervised learning [91], among which CRFs have often been used either in pre-processing pseudo ground truth labels or post-processing results from DL models trained on weak labels. In our case, we designed the system using a fully connected CRF model as a post-processing step, which further improved the accuracy of the initial automatic segmentation based on the spatial and intensity consistency of the aneurysms. Some other works have also reported the application of a subsequent refinement neural network [92], but this approach often requires a small number of refined manual segmentations and will lead to a semi-supervised

framework instead.

5.6 Conclusion

In conclusion, we have proposed a novel 3D focal modulation UNet called FocalSegNet in combination with CRF for weakly supervised segmentation and detection of brain aneurysms from TOF-MRA. By leveraging coarse segmentation ground truths, the proposed technique was able to achieve excellent performance. By comparing its performance with the popular Residual-UNet and the most recent Swin-UNETR, we demonstrated its superior performance and will extend its application to other domain tasks in the near future.

Chapter 6

Conclusion and Future Work

In this thesis, we explored the application of weakly supervised deep learning methods for cerebrovascular pathology segmentation, with three novel methods. In our first method, we explored the use of self-attention maps to segment intracranial hemorrhages (ICH). Two Swin transformer models were trained based on categorical labels for binary ICH detection and full ICH subtyping. Subsequently, we developed a framework that combines attention maps from different layers of these Swin transformers in a hierarchical manner to produce a hemorrhage localization map. This map was then subjected to thresholding to generate a discrete segmentation map. Our experiments demonstrated that training the Swin transformer with binary labels (with/without ICH) enhances the concentration of attention maps on relevant hemorrhagic regions. We also compared three thresholding techniques - k-means, Otsu's method, and simple thresholding - and found that the simple thresholding is the most advantageous.

With these insights, we further introduced a more robust framework called head-wise gradient-infused self-attention mapping (HGI-SAM) to perform weakly supervised ICH segmentation. HGI-SAM utilizes a Swin transformer model for binary ICH detection and continue to use simple thresholding to obtain discrete segmentation maps. Specifically, it improves the quality of hemorrhage localization map by performing head-wise weighing of self-attention obtained from the Swin transformer using the gradient of the target class. This approach resulted in enhanced segmentation accuracy and interpretability of the Swin transformer, with similar performance in comparison with popular fully supervised DL segmentation methods, thus highlighting the strength of our proposed

framework.

In the last contribution of the thesis, we shifted our focus to the segmentation of unruptured intracranial aneurysms, another cerebrovascular disorder that may lead to ICH, and explored a different weakly supervised learning technique that involves coarse segmentations. Specifically, we proposed a novel 3D focal modulation UNet, named FocalSegNet that was trained with rough contouring of the aneurysms as segmenation ground truths, in conjunction with conditional random field post-processing to enable refined segmentation of the pathology. We conducted comprehensive ablation studies on our design choices and compared the performance of the newly proposed FocalSegNet against the popular UNet and Swin-UNETR model to confirm the superiority of the proposed framework.

Looking ahead, a few perspectives of the presented methods can be further extended. For the case of weakly supervised ICH segmenation, our proposed methods are based on 2D CT slice processing while leveraging the inter-slice or 3D anatomical context could potentially boost the accuracy of the methods. However, to do so, additional exploration is required to allow accomodation of the variations in image slice thickness from clinical CT scans. To obtain the final segmentation, we have opted to use simple thresholding. Future studies could also be directed to the investigation of extra light-weight refinement networks and iterative training schemes to furhter improve the quality of segmentation. For aneurysm segmentation, we have demonstrated the advantage of focal modulation when being trained on coarse image segmentation. Instead of using conditional random field as a post-processing step, one potential future direction for our work could be exploring relevant regularization functions to guide the training process that allows direct production of refined segmentation results. Finally, in addition to weakly supervised techniques, self-supervised and unsupervised learning methods have also emerged recently as exciting avenues to help mitigate the limited training data in medical deep learning. These methods may also greatly contribute to the care of cerebravascular care, which emphasizes both accuracy and fast uptake.

Bibliography

[1] J. Frösen, R. Tulamo, A. Paetau, E. Laaksamo, M. Korja, A. Laakso, M. Niemelä, and J. Hernesniemi, "Saccular intracranial aneurysm: pathology and mechanisms," *Acta Neuropathol*, vol. 123, no. 6, pp. 773–86, 2012.

[2] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," *Computer methods in biomechanics and biomedical engineering*, vol. 17, no. 2, pp. 73–93, 2014.

[3] Z. Hou *et al.*, "A review on mr image intensity inhomogeneity correction," *International Journal of Biomedical Imaging*, 2006.

[4] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[5] J. Nalepa, M. Marcinkiewicz, and M. Kawulok, "Data augmentation for brain-tumor segmentation: a review," *Frontiers in computational neuroscience*, vol. 13, p. 83, 2019.

[6] A. Xu, L. Wang, S. Feng, and Y. Qu, "Threshold-based level set method of image segmentation," in *2010 Third International Conference on Intelligent Networks and Intelligent Systems*, pp. 703–706, IEEE, 2010.

[7] C. Cigla and A. A. Alatan, "Region-based image segmentation via graph cuts," in *2008 15th IEEE International Conference on Image Processing*, pp. 2272–2275, IEEE, 2008.

[8] Z. Yu-Qian, G. Wei-Hua, C. Zhen-Cheng, T. Jing-Tian, and L. Ling-Yun, "Medical images

edge detection based on mathematical morphology," in *2005 IEEE engineering in medicine and biology 27th annual conference*, pp. 6492–6495, IEEE, 2006.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[11] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

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

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[16] J. Yang, C. Li, X. Dai, L. Yuan, and J. Gao, "Focal modulation networks," *arXiv preprint arXiv:2203.11926*, 2022.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected

crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[18] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans Med Imaging*, vol. 36, no. 2, pp. 674–683, 2017.

[19] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," *arXiv preprint arXiv:2004.06816*, 2020.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[22] O. Viniavskyi, M. Dobko, and O. Dobosevych, "Weakly-supervised segmentation for disease localization in chest x-ray images," in *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pp. 249–259, Springer, 2020.

[23] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2209–2218, 2019.

[24] H. Salehinejad, J. Kitamura, N. Ditkofsky, A. Lin, A. Bharatha, S. Suthiphosuwan, H.-M. Lin, J. R. Wilson, M. Mamdani, and E. Colak, "A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography," *Scientific Reports*, vol. 11, no. 17051, 2021.

[25] D. Rajashekar and J. W. Liang, "Intracerebral hemorrhage," in *StatPearls [Internet]*, StatPearls Publishing, 2021.

[26] T. Apostolaki-Hansson, T. Ullberg, M. Pihlsgård, B. Norrving, and J. Petersson, "Prognosis of intracerebral hemorrhage related to antithrombotic use: an observational study from the swedish stroke register (riksstroke)," *Stroke*, vol. 52, no. 3, pp. 966–974, 2021.

[27] A. Qureshi and Y. Palesch, "Antihypertensive treatment of acute cerebral hemorrhage (atach) ii: design, methods, and rationale," *Neurocritical Care*, vol. 15, no. 3, pp. 559–576, 2011.

[28] M. D. Hssayeni, M. S. Croock, A. D. Salman, H. F. Al-khafaji, Z. A. Yahya, and B. Ghoraani, "Intracranial hemorrhage segmentation using a deep convolutional model," *Data*, vol. 5, no. 1, p. 14, 2020.

[29] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.

[30] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: Residual vision transformers for multi-modal medical image synthesis," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2022.

[31] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[32] J. Cho, K.-S. Park, M. Karki, E. Lee, S. Ko, J. K. Kim, D.-E. Lee, J. Choe, J. Son, M. Kim, S. Lee, J.-E. Lee, C. H. Yoon, and S. youl Park, "Improving sensitivity on identification and delineation of intracranial hemorrhage lesion using cascaded deep learning models," *Journal of Digital Imaging*, vol. 32, pp. 450–461, 2018.

[33] H. Ye, F. Gao, Y. Yin, D. Guo, P. Zhao, Y. Lu, X. Wang, J. Bai, K. Cao, Q. Song, H. Zhang, W. Chen, X. Guo, and J. Xia, "Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network," *European Radiology*, vol. 29, pp. 6191–6201, 2019.

[34] K. Wu, B. Du, M. Luo, H. Wen, Y. Shen, and J. Feng, "Weakly supervised brain lesion segmentation via attentional representation learning," in *Medical Image Computing and Computer*

Assisted Intervention - MICCAI 2019, (Cham), pp. 211–219, Springer International Publishing, 2019.

[35] J. Nemcek, T. Vicar, and R. Jakubicek, "Weakly supervised deep learning-based intracranial hemorrhage localization," *arXiv preprint arXiv:2105.00781*, 2021.

[36] R. Wightman, "Pytorch image models." https://github.com/rwightman/pytorch-image-models, 2019.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[38] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

[39] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

[40] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018, p. 2225, NIH Public Access, 2018.

[41] V. A. Sindagi and V. M. Patel, "Ha-ccn: Hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2019.

[42] A. E. Flanders, L. M. Prevedello, G. Shih, S. S. Halabi, J. Kalpathy-Cramer, R. Ball, J. T. Mongan, A. Stein, F. C. Kitamura, M. P. Lungren, *et al.*, "Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge," *Radiology: Artificial Intelligence*, vol. 2, no. 3, 2020.

[43] D. Alis, C. Alis, M. Yergin, C. Topel, O. Asmakutlu, O. Bagcilar, Y. D. Senli, A. Ustundag, V. Salt, S. N. Dogan, *et al.*, "A joint convolutional-recurrent neural network with an attention mechanism for detecting intracranial hemorrhage on noncontrast head ct," *Scientific Reports*, vol. 12, no. 1, p. 2084, 2022.

[44] A. Rasoulian, S. Salari, and Y. Xiao, "Weakly supervised intracranial hemorrhage segmentation using hierarchical combination of attention maps from a swin transformer," in *Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pp. 63–72, Springer, 2022.

[45] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6202–6212, 2023.

[46] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pp. 205–218, Springer, 2023.

[47] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pp. 272–284, Springer, 2022.

[48] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[49] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[50] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, and P. Szczuko, "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Information Fusion*, vol. 90, pp. 316–352, 2023.

[51] S. Syed, K. E. Anderssen, S. K. Stormo, and M. Kranz, "Weakly supervised semantic segmentation for mri: exploring the advantages and disadvantages of class activation maps for biological image segmentation with soft boundaries," *Scientific Reports*, vol. 13, no. 1, p. 2574, 2023.

[52] M. Yurt, O. Dalmaz, S. Dar, M. Ozbey, B. Tınaz, K. Oguz, and T. Çukur, "Semi-supervised learning of mri synthesis without fully-sampled ground truths," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3895–3906, 2022.

[53] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. Van Tulder, and M. De Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pp. 457–465, Springer, 2019.

[54] Y. Zhou, X. He, L. Huang, L. Liu, F. Zhu, S. Cui, and L. Shao, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2079–2088, 2019.

[55] X. Liu, Q. Yuan, Y. Gao, K. He, S. Wang, X. Tang, J. Tang, and D. Shen, "Weakly supervised segmentation of covid19 infection with scribble annotation on ct images," *Pattern Recognition*, vol. 122, p. 108341, 2022.

[56] H. R. Roth, D. Yang, Z. Xu, X. Wang, and D. Xu, "Going to extremes: Weakly supervised medical image segmentation," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 507–524, 2021.

[57] H. Lin, H. Chen, Q. Dou, L. Wang, J. Qin, and P.-A. Heng, "Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 539–546, IEEE, 2018.

[58] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[59] Y. Han, L. Cheng, G. Huang, G. Zhong, J. Li, X. Yuan, H. Liu, J. Li, J. Zhou, and M. Cai, "Weakly supervised semantic segmentation of histological tissue via attention accumulation and pixel-level contrast learning," *Physics in Medicine and Biology*, 2022.

[60] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-cam: Causal cam for weakly supervised semantic segmentation on medical image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11676–11685, 2022.

[61] M. Yu, M. Han, X. Li, X. Wei, H. Jiang, H. Chen, and R. Yu, "Adaptive soft erasure with edge self-attention for weakly supervised semantic segmentation: thyroid ultrasound image case study," *Computers in Biology and Medicine*, vol. 144, p. 105347, 2022.

[62] K. Li, Z. Qian, Y. Han, I. Eric, C. Chang, B. Wei, M. Lai, J. Liao, Y. Fan, and Y. Xu, "Weakly supervised histopathology image segmentation with self-attention," *Medical Image Analysis*, p. 102791, 2023.

[63] S. Zhang, J. Zhang, and Y. Xia, "Transws: Transformer-based weakly supervised histology image segmentation," in *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pp. 367–376, Springer, 2022.

[64] K. Saab, J. Dunnmon, R. Goldman, A. Ratner, H. Sagreiya, C. Ré, and D. Rubin, "Doubly weak supervision of deep learning models for head ct," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pp. 811–819, Springer, 2019.

[65] Y. Liu, E. Zhang, L. Xu, C. Xiao, X. Zhong, L. Lian, F. Li, B. Jiang, Y. Dong, L. Ma, *et al.*, "Mixed-unet: Refined class activation mapping for weakly-supervised semantic segmentation with multi-scale inference," *arXiv preprint arXiv:2205.04227*, 2022.

[66] W. Sun, J. Zhang, Z. Liu, Y. Zhong, and N. Barnes, "Getam: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation," *arXiv preprint arXiv:2112.02841*, 2021.

[67] O. Barkan, E. Hauon, A. Caciularu, O. Katz, I. Malkiel, O. Armstrong, and N. Koenigstein, "Grad-sam: Explaining transformers via gradient self-attention maps," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2882–2887, 2021.

[68] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, "Ts-cam: Token semantic coupled attention map for weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2886–2895, 2021.

[69] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.

[70] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *ArXiv*, vol. abs/2103.11886, 2021.

[71] J. Muschelli, "Recommendations for processing head ct data," *Frontiers in neuroinformatics*, vol. 13, p. 61, 2019.

[72] J. Gildenblat and contributors, "Pytorch library for cam methods." https://github.com/jacobgil/pytorch-grad-cam, 2021.

[73] K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, X. An, M. Baumgartner, P. Bourdon, R. De Feo, T. Di Noto, F. Dubost, *et al.*, "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge," *Neuroimage*, vol. 238, p. 118216, 2021.

[74] Y. Liang, M. Li, and C. Jiang, "Generating self-attention activation maps for visual interpretations of convolutional neural networks," *Neurocomputing*, vol. 490, pp. 206–216, 2022.

[75] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102789, 2023.

[76] Y. Sui, O. Afacan, A. Gholipour, and S. K. Warfield, "Mri super-resolution through generative degradation learning," *Med Image Comput Comput Assist Interv*, vol. 12906, pp. 430–440, 2021.

[77] P. M. White, J. M. Wardlaw, and V. Easton, "Can noninvasive imaging accurately depict intracranial aneurysms? a systematic review," *Radiology*, vol. 217, no. 2, pp. 361–70, 2000.

[78] K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, X. An, M. Baumgartner, P. Bourdon, R. De Feo, T. D. Noto, F. Dubost, A. Fava-Sanches, X. Feng, C. Giroud, I. Group, M. Hu, P. F. Jaeger, J. Kaiponen, M. Klimont, Y. Li, H. Li, Y. Lin, T. Loehr, J. Ma, K. H. Maier-Hein, G. Marie, B. Menze, J. Richiardi, S. Rjiba, D. Shah, S. Shit, J. Tohka, T. Urruty, U. Walińska, X. Yang, Y. Yang, Y. Yin, B. K. Velthuis, and H. J. Kuijf, "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge," *Neuroimage*, vol. 238, p. 118216, 2021.

[79] M. Din, S. Agarwal, M. Grzeda, D. A. Wood, M. Modat, and T. C. Booth, "Detection of cerebral aneurysms using artificial intelligence: a systematic review and meta-analysis," *J Neurointerv Surg*, vol. 15, no. 3, pp. 262–271, 2023.

[80] G. Yang, C. Wang, J. Yang, Y. Chen, L. Tang, P. Shao, J.-L. Dillenseger, H. Shu, and L. Luo, "Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal cta images," *BMC Medical Imaging*, vol. 20, no. 1, p. 37, 2020.

[81] A. Rasoulian, S. Salari, and Y. Xiao, "Weakly supervised intracranial hemorrhage segmentation using head-wise gradient-infused self-attention maps from a swin transformer in categorical learning," *Machine Learning for Biomedical Imaging*, vol. 2, pp. 338–360, 2023.

[82] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, p. 424–432, 2016.

[83] M. Naderi, M. Givkashi, F. Piri, N. Karimi, and S. Samavi, "Focal-unet: Unet-like focal modulation for medical image segmentation," *arXiv preprint arXiv:2212.09263*, 2022.

[84] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, 2011.

[85] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, (San Francisco, CA, USA), pp. 282–289, Morgan Kaufmann Publishers Inc., 2001.

[86] W. Silversmith, "cc3d: Connected components on multilabel 3D and 2D images.," sep 2021.

[87] T. Di Noto, G. Marie, S. Tourbier, Y. Alemán-Gómez, O. Esteban, G. Saliou, M. B. Cuadra, P. Hagmann, and J. Richiardi, "Towards automated brain aneurysm detection in tof-mra: Open data, weak labels, and anatomical knowledge," *Neuroinformatics*, vol. 21, no. 1, pp. 21–34, 2023.

[88] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," *IEEE Trans Med Imaging*, vol. 29, no. 6, pp. 1310–20, 2010.

[89] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pp. 240–248, Springer, 2017.

[90] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," *Medical image analysis*, vol. 67, p. 101851, 2018.

[91] L. Chan, M. S. Hosseini, and K. N. Plataniotis, "A comprehensive analysis of weakly-supervised semantic segmentation in different image domains," *International Journal of Computer Vision*, vol. 129, pp. 361–384, 2021.

[92] S. Adiga V, J. Dolz, and H. Lombaert, "Manifold-driven attention maps for weakly supervised segmentation," *arXiv preprint arXiv:2004.03046*, 2020.