Weakly Supervised Intracranial Hemorrhage Segmentation using Head-Wise Gradient-Infused Self-Attention Maps from a Swin Transformer in Categorical Learning

Amirhossein Rasoulian https://orcid.org/0009-0004-9077-521X   ah.rasoulian@gmail.com
Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada

Soorena Salari https://orcid.org/0000-0002-2587-0323   soorena.salari374@gmail.com
Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada

Yiming Xiao https://orcid.org/0000-0002-0962-3525   yiming.xiao@concordia.ca
Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada

Abstract

Intracranial hemorrhage (ICH) is a life-threatening medical emergency that requires timely and accurate diagnosis for effective treatment and improved patient survival rates. While deep learning 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 facilitate ICH treatment decisions, we introduce a novel weakly supervised method for ICH segmentation, utilizing a Swin transformer trained on an ICH classification task with categorical labels. Our approach leverages a hierarchical combination of head-wise gradient-infused self-attention maps to generate accurate image segmentation. Additionally, we conducted an exploratory study on different learning strategies and showed that binary ICH classification has a more positive impact on self-attention maps compared to full ICH subtyping. With a mean Dice score of 0.44, our technique achieved similar ICH segmentation performance as the popular U-Net and Swin-UNETR models with full supervision and outperformed a similar weakly supervised approach using GradCAM, demonstrating the excellent potential of the proposed framework in challenging medical image segmentation tasks. Our code is available at https://github.com/HealthX-Lab/HGI-SAM.

Keywords: Weak Supervision, Image Segmentation, Swin Transformer, Intracranial Hemorrhage, Self-attention

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 (Rajashekar and Liang, 2021; Apostolaki-Hansson et al., 2021). The outcome of ICH depends on the volume of bleeding, which can enlarge rapidly within the first few hours (Qureshi and Palesch, 2011), 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 (Hssayeni et al., 2020; Alis et al., 2022; Salehinejad et al., 2021). 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) (Dosovitskiy et al., 2021) 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 (Dai et al., 2021; Dalmaz et al., 2022). 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 (Liu et al., 2021) 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 (Zhou, 2017) 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 (Rasoulian et al., 2022), 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 (Rasoulian et al., 2022) with three main contributions. **First**, inspired by the gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017), 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 (Rasoulian et al., 2022). **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 (Flanders et al., 2020) and PhysioNet datasets (Hssayeni et al., 2020), 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.

2. Related Works

There have been several variants of the Swin transformer model for medical image segmentation tasks. Heidari et al. (2023) 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. (2023) proposed a Swin-based U-Net-like model to segment abdominal CT images and cardiac MRI scans. Hatamizadeh et al. (2022) proposed a hybrid Swin-encoder-CNN-decoder model to segment brain tumor MRI images. Finally, Lin et al. (2022) 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 (Wang et al., 2022; Qureshi et al., 2023; Syed et al., 2023). 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. (2022) used Generative Adversarial Networks (GANs) for MRI contrast translation with undersampled k-space data. Chen et al. (2019) employed attention-based multi-task learning that simultaneously optimizes a supervised segmentation and an unsupervised reconstruction for brain tumor segmentation. Finally, Zhou et al. (2019) 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 (Rajchl et al., 2017), scribbles (Liu et al., 2022a), points (Roth et al., 2021), or even categorical labels (Lin et al., 2018). 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. (2017) proposed DeepCut, an ap-

proach 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. (2020) employed global constraints derived from box annotations, including tightness prior and global background emptiness, to achieve improved segmentation results over DeepCut (Rajchl et al., 2017) on the PROMISE12 dataset (Litjens et al., 2014). Previously, scribble and point annotations have been widely used in interactive segmentation. In weakly supervised segmentation, Roth et al. (2021) 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. (2022a) 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. (2022) proposed a weakly supervised segmentation model based on class residual attention for the lung adenocarcinoma and breast cancer datasets. Chen et al. (2022) 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. (2020) utilized class activation maps (CAM) for Chest X-Ray segmentation. More recently, Yu et al. (2022) further modified CAMs by scale feature adaptation and soft-erase modules to segment thyroid ultrasound images. With the transformer model, Li et al. (2023) utilized a self-attention mechanism in multiple instances learning for weakly supervised segmentation of histopathology images while Zhang et al. (2022) 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. (2020) 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 0.846~0.975, while more fine-grained ICH subtyping achieved an AUC of 0.93~0.96. Deep learning-based approaches in ICH detection typically used fully convolutional networks (FCNs) (Cho et al., 2018) and recurrent neural networks (RNNs) (Ye et al., 2019), 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. (2019) and Salehinejad et al. (2021) 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 (Wu et al., 2019; Nemcek et al., 2021; Liu et al., 2022b). Specifically, Wu et al. (2019) 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
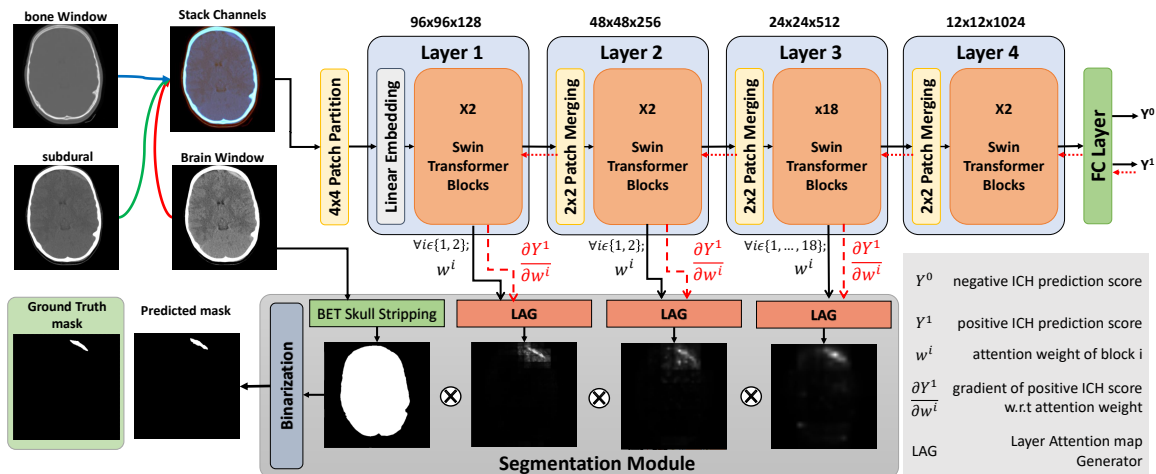
Figure 1: Overview of the proposed weakly supervised segmentation method using hierarchical fusion of gradient-weighted self-attention maps.

et al. (2022b) 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. (2021) 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 (Rasoulian et al., 2022), 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.

3. Methods

An overview of our proposed weakly supervised technique for ICH segmentation is depicted in Fig. 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.

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. 2a). The self-attention per head within each window is computed as:

$$
\begin{aligned}
\text{Attention}\,(Q_h, K_h, V_h) &= w_h \times V_h, \\
w_h &= \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d}} + B_h\right),
\end{aligned}
\tag{1}
$$

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 (Liu et al., 2021).

In our earlier study (Rasoulian et al., 2022), 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.

3.2 Hemorrhage Segmentation

In our previous study (Rasoulian et al., 2022), 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 (Chefer et al., 2021; Dosovitskiy et al., 2021). 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 regu-
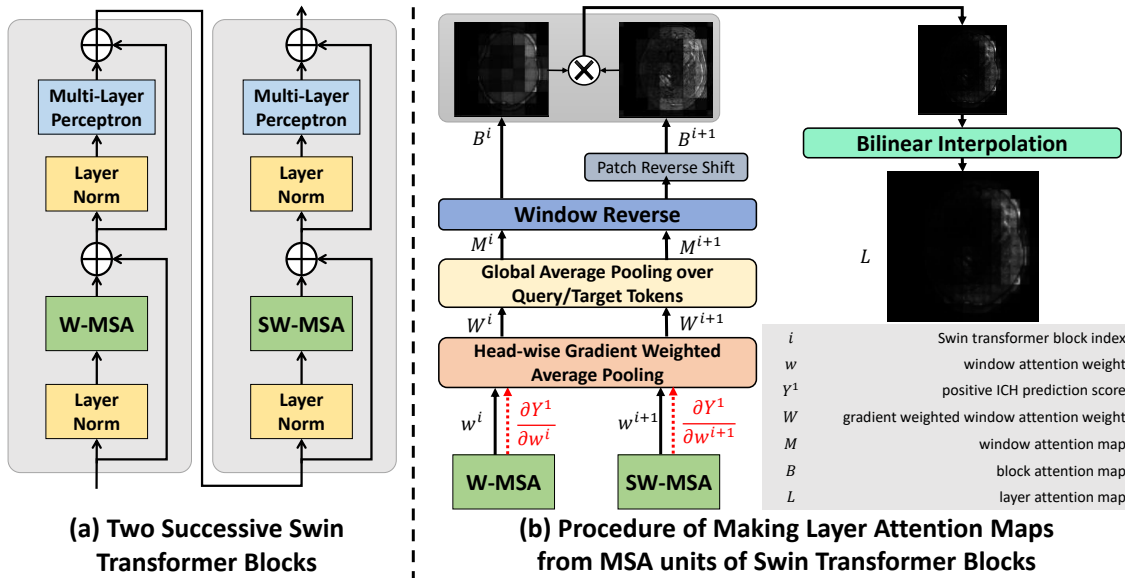
Figure 2: Demonstration of head-wise gradient-infused layer attention map generation in the proposed Swin transformer in categorical learning

lar 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.

3.2.1 LAYER ATTENTION MAP GENERATION

There has been recent research that leverages model classification scores for attention explainability. For instance, Chefer et al. (2021) utilize the Taylor Decomposition principle to assign and propagate a local relevance score through the layers of a ViT model. Similarly, Sun et al. (2021) and Barkan et al. (2021) 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 (Gao et al., 2021), including our own previous work (Rasoulian et al., 2022). 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. (2019), 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. 2, the layer attention map is created as follows:

$$
\begin{aligned}
W^i &= \frac{1}{H} \sum_{h=1}^{H} \left\| \frac{\partial Y^1}{\partial w^i{}_h} \right\| \cdot w^i{}_h, \\
M^i &= \frac{1}{Q} \sum_{k=1}^{Q} W_k^i, \\
L &= \mathrm{BI} \left(\mathrm{WR} \left(W^i \right) \otimes \mathrm{RS} \left(\mathrm{WR}(W^{i+1}) \right) \right)
\end{aligned}
\tag{2}
$$

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.

3.2.2 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 (Rasoulian et al., 2022), 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 (Zhou et al., 2021). Furthermore, as demonstrated in Fig. 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
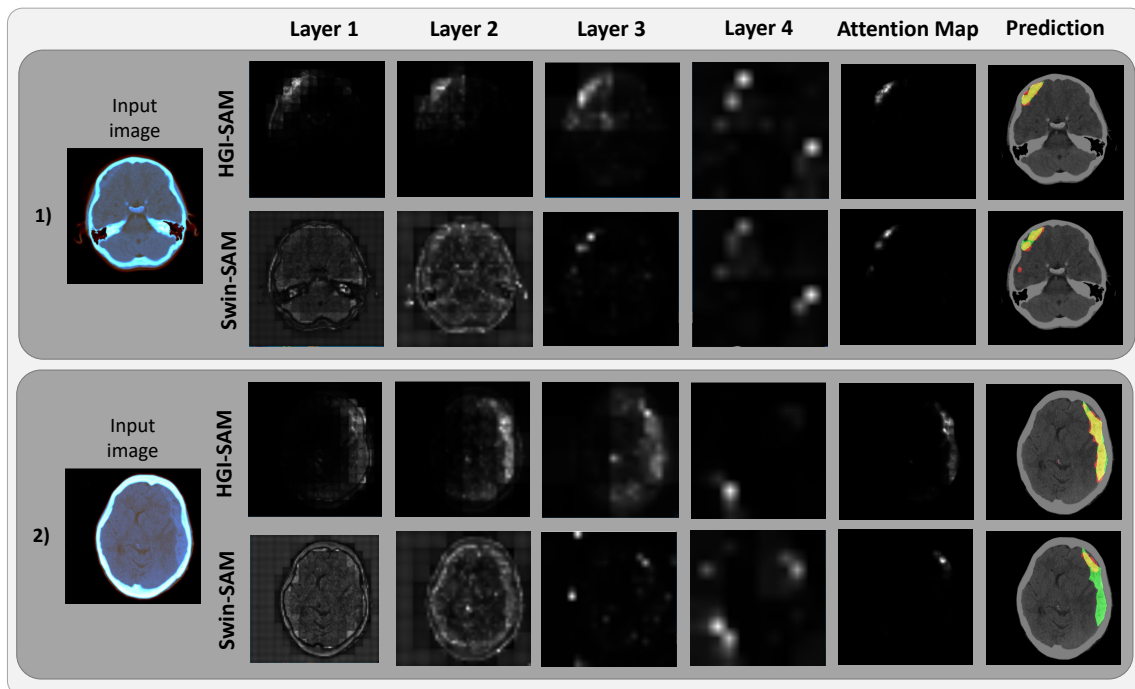
Figure 3: A comparison of the proposed head-wise gradient-infused self-attention mapping (HGI-SAM) and the original self-attention maps (Swin-SAM) from the Swin transformer model is shown for two axial CT slices. Along with the maps at different hierarchies, the fused attention maps and the derived binary 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).

(2019). 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 (Rasoulian et al., 2022), 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. 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 (Rasoulian et al., 2022), 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.

4.1 Dataset

To train and evaluate our models, we used two public datasets, the RSNA ICH CT dataset (Flanders et al., 2020) and the PhysioNet CT dataset (Hssayeni et al., 2020). 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 (Hatamizadeh et al., 2022), 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 (Flanders et al., 2020; Hssayeni et al., 2020) 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].

4.2 Implementation details

4.2.1 ICH SEGMENTATION WITH SWIN-SAM MULTI-LABEL AND SWIN-SAM BINARY

In our previous work (Rasoulian et al., 2022), two Swin transformers (Wightman, 2019) 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.

4.2.2 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 (Rasoulian et al., 2022). 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.

4.3 Baseline models

4.3.1 ICH segmentation with Grad-CAM

As most existing weakly supervised segmentation techniques relied on Grad-CAM (Selvaraju et al., 2017), 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 Gildenblat and contributors (2021), 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.

4.3.2 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 (Hssayeni et al., 2020), 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.

4.3.3 Fully supervised Swin-UNETR

Swin-UNETR (Hatamizadeh et al., 2022) 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 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 (Timmins et al., 2021). 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.

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. 3 for the axial CT slices of two patients. From Fig. 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. 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 (Rasoulian et al., 2022), 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 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 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 1 and Table 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.

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 (Selvaraju et al., 2017) 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 (Chen et al., 2022), including applications for stroke lesion segmentation (Wu et al., 2019; Nemcek et al., 2021). Compared with Grad-CAM and its variants, the more recent attention mechanisms, especially self-attention from transformer models, can identify more discriminative, task-related regions and features while improving the performance of the DL models (Liang et al., 2022; Dosovitskiy et al., 2021). 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

Table 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	$\mathbf{0.237 \pm 0.012}$	$\mathbf{0.299 \pm 0.012}$	$\mathbf{0.387 \pm 0.012}$	$\mathbf{0.444 \pm 0.014}$	$\mathbf{0.438 \pm 0.019}$	$\mathbf{0.455 \pm 0.019}$

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	$\mathbf{0.151 \pm 0.008}$	$\mathbf{0.196 \pm 0.009}$	$\mathbf{0.263 \pm 0.010}$	$\mathbf{0.319 \pm 0.012}$	$\mathbf{0.343 \pm 0.016}$	$\mathbf{0.355 \pm 0.016}$

Table 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**
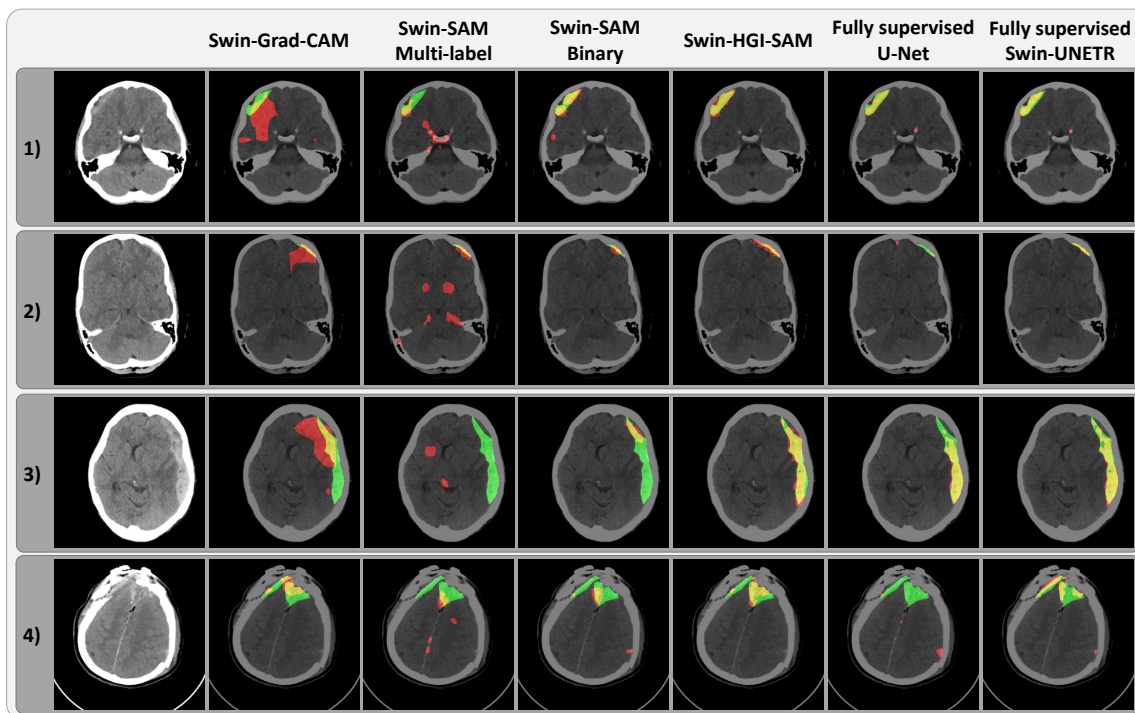
Figure 4: Qualitative comparison of segmentation performance for the proposed weakly supervised ICH segmentation methods (Swin-Grad-CAM, Swin-SAM Multi-label, Swin-SAM Binary, and Swin-HGI-SAM), fully supervised U-Net, and fully supervised Swin-UNETR for four different cases. Here, red=automatic segmentation, green=ground truths, and yellow=true positives.

the downstream segmentation outcomes in the proposed framework. As an ablation study of our previous work (Rasoulian et al., 2022), 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 (Selvaraju et al., 2017), 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 (Chefer et al., 2021; Sun et al., 2021) 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. 3 and Fig. 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 (Zhou et al., 2021). 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 (Rasoulian et al., 2022), 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 (Hssayeni et al., 2020). 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 (Billot et al., 2023) and image super-resolution (Sui et al., 2021) 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.

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.

Acknowledgments

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

The authors declare no conflicts of interest for this work.

References

Deniz Alis, Ceren Alis, Mert Yergin, Cagdas Topel, Ozan Asmakutlu, Omer Bagcilar, Yeseren Deniz Senli, Ahmet Ustundag, Vefa Salt, Sebahat Nacar Dogan, et al. A joint convolutional-recurrent neural network with an attention mechanism for detecting intracranial hemorrhage on noncontrast head ct. *Scientific Reports*, 12(1):2084, 2022.

Trine Apostolaki-Hansson, Teresa Ullberg, Mats Pihlsgård, Bo Norrving, and Jesper Petersson. Prognosis of intracerebral hemorrhage related to antithrombotic use: an observational study from the swedish stroke register (riksstroke). *Stroke*, 52(3):966–974, 2021.

Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2882–2887, 2021.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. ISSN 1361-8415.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning 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*, pages 205–218. Springer, 2023.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs Van Tulder, and Marleen 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*, pages 457–465. Springer, 2019.

Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi 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*, pages 11676–11685, 2022.

Junghwan Cho, Ki-Su Park, Manohar Karki, Eunmi Lee, Seokhwan Ko, Jong Kun Kim, Dong-Eun Lee, Jaeyoung Choe, Jeongwoo Son, Myungsoo Kim, Sukhee Lee, Jeong-Eom Lee, Chang Hyo Yoon, and Sin youl Park. Improving sensitivity on identification and delineation of intracranial hemorrhage lesion using cascaded deep learning models. *Journal of Digital Imaging*, 32:450–461, 2018.

Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021. ISSN 2075-4418.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multi-modal medical image synthesis. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2 (3), 2020.

Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021.

Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

Yongqi Han, Lianglun Cheng, Guoheng Huang, Guo Zhong, Jiahua Li, Xiaochen Yuan, Hongrui Liu, Jiao Li, Jian Zhou, and Muyan Cai. Weakly supervised semantic segmentation of histological tissue via attention accumulation and pixel-level contrast learning. *Physics in Medicine and Biology*, 2022.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang 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*, pages 272–284. Springer, 2022.

Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit 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*, pages 6202–6212, 2023.

Murtadha D Hssayeni, Muayad S Croock, Aymen D Salman, Hassan Falah Al-khafaji, Zakaria A Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1):14, 2020.

Hoel Kervadec, José Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *ArXiv*, abs/2004.06816, 2020.

Kailu Li, Ziniu Qian, Yingnan Han, I Eric, Chao Chang, Bingzheng Wei, Maode Lai, Jing Liao, Yubo Fan, and Yan Xu. Weakly supervised histopathology image segmentation with self-attention. *Medical Image Analysis*, page 102791, 2023.

Yu Liang, Maozhen Li, and Changjun Jiang. Generating self-attention activation maps for visual interpretations of convolutional neural networks. *Neurocomputing*, 490:206–216, 2022. ISSN 0925-2312.

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

Huangjing Lin, Hao Chen, Qi Dou, Liansheng Wang, Jing Qin, and Pheng-Ann 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)*, pages 539–546. IEEE, 2018.

Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

Xiaoming Liu, Quan Yuan, Yaozong Gao, Kelei He, Shuo Wang, Xiao Tang, Jinshan Tang, and Dinggang Shen. Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. *Pattern Recognition*, 122:108341, 2022a. ISSN 0031-3203.

Yang Liu, Ersi Zhang, Lulu Xu, Chufan Xiao, Xiaoyun Zhong, Lijin Lian, Fang Li, Bin Jiang, Yuhan Dong, Lan Ma, et al. Mixed-unet: Refined class activation mapping for weakly-supervised semantic segmentation with multi-scale inference. *arXiv preprint arXiv:2205.04227*, 2022b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

John Muschelli. Recommendations for processing head ct data. *Frontiers in neuroinformatics*, 13:61, 2019.

Jakub Nemcek, Tomas Vicar, and Roman Jakubicek. Weakly supervised deep learning-based intracranial hemorrhage localization. *arXiv preprint arXiv:2105.00781*, 2021.

AI Qureshi and YY Palesch. Antihypertensive treatment of acute cerebral hemorrhage (atach) ii: design, methods, and rationale. *Neurocritical Care*, 15(3):559–576, 2011.

Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid, Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352, 2023. ISSN 1566-2535.

Devika Rajashekar and John W Liang. Intracerebral hemorrhage. In *StatPearls [Internet]*. StatPearls Publishing, 2021.

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*, 36(2):674–683, 2017. ISSN 1558-254X (Electronic) 0278-0062 (Print) 0278-0062 (Linking).

Amirhossein Rasoulian, Soorena Salari, and Yiming 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*, pages 63–72. Springer, 2022. ISBN 978-3-031-17898-6.

Holger R. Roth, Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Going to extremes: Weakly supervised medical image segmentation. *Machine Learning and Knowledge Extraction*, 3(2):507–524, 2021.

Khaled Saab, Jared Dunnmon, Roger Goldman, Alex Ratner, Hersh Sagreiya, Christopher Ré, and Daniel 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*, pages 811–819. Springer, 2019.

Hojjat Salehinejad, Jumpei Kitamura, Noah Ditkofsky, Amy Lin, Aditya Bharatha, Suradech Suthiphosuwan, Hui-Ming Lin, Jefferson R Wilson, Muhammad Mamdani, and Errol Colak. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. *Scientific Reports*, 11(17051), 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

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

Weixuan Sun, Jing Zhang, Zheyuan Liu, Yiran Zhong, and Nick Barnes. Getam: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.02841*, 2021.

Shaheen Syed, Kathryn E Anderssen, Svein Kristian Stormo, and Mathias 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*, 13(1):2574, 2023.

Kimberley M Timmins, Irene C van der Schaaf, Edwin Bennink, Ynte M Ruigrok, Xingle An, Michael Baumgartner, Pascal Bourdon, Riccardo De Feo, Tommaso Di Noto, Florian Dubost, et al. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge. *Neuroimage*, 238:118216, 2021.

Ostap Viniavskyi, Mariia Dobko, and Oles 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*, pages 249–259. Springer, 2020.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16 (5):1243–1267, 2022.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Kai Wu, Bowen Du, Man Luo, Hongkai Wen, Yiran Shen, and Jianfeng Feng. Weakly supervised brain lesion segmentation via attentional representation learning. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, pages 211–219, Cham, 2019. Springer International Publishing.

Hai Ye, Feng Gao, Youbing Yin, Danfeng Guo, Pengfei Zhao, Yi Lu, Xin Wang, Junjie Bai, Kunlin Cao, Qi Song, Heye Zhang, Wei Chen, Xuejun Guo, and Jun Xia. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European Radiology*, 29:6191–6201, 2019.

Mei Yu, Ming Han, Xuewei Li, Xi Wei, Han Jiang, Huiling Chen, and Ruiguo Yu. Adaptive soft erasure with edge self-attention for weakly supervised semantic segmentation: thyroid ultrasound image case study. *Computers in Biology and Medicine*, 144:105347, 2022.

Mahmut Yurt, Onat Dalmaz, Salman Dar, Muzaffer Ozbey, Berk Tınaz, Kader Oguz, and Tolga Çukur. Semi-supervised learning of mri synthesis without fully-sampled ground truths. *IEEE Transactions on Medical Imaging*, 41(12):3895–3906, 2022.

Shaoteng Zhang, Jianpeng Zhang, and Yong 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*, pages 367–376. Springer, 2022.

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *ArXiv*, abs/2103.11886, 2021.

Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling 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*, pages 2079–2088, 2019.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.