Localization-Guided Supervision for Robust Medical Image Classification by Vision Transformers

Sagi Ben Itzhak[1], Nahum Kiryati[2], Orith Portnoy[3], and Arnaldo Mayer[3]

[1] School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel
sagib2@mail.tau.ac.il
[2] Klachky Chair of Image Processing, School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel
nk@eng.tau.ac.il
[3] Diagnostic Imaging Department at Sheba Medical Center Affiliated with the School of Medicine, Tel Aviv University, Tel Aviv, Israel
arnaldo.mayer@sheba.health.gov.il

Abstract. A major challenge in developing data-driven algorithms for medical imaging is the limited size of available datasets. Furthermore, these datasets often suffer from inter-site heterogeneity caused by the use of different scanners and scanning protocols. These factors may contribute to overfitting, which undermines the generalization ability and robustness of deep learning classification models in the medical domain, leading to inadequate performance in real-world applications. To address these challenges and mitigate overfitting, we propose a framework which incorporates explanation supervision during training of Vision Transformer (ViT) models for image classification. Our approach leverages foreground masks of the class object during training to regularize attribution maps extracted from ViT, encouraging the model to focus on relevant image regions and make predictions based on pertinent features. We introduce a new method for generating explanatory attribution maps from ViT-based models and construct a dual-loss function that combines a conventional classification loss with a term that regularizes attribution maps. Our approach demonstrates superior performance over existing methods on two challenging medical imaging datasets, highlighting its effectiveness in the medical domain and its potential for application in other fields. Source code is available at: https://github.com/sagibe/LGMViT

Keywords: Explainability · Explainable AI · Vision Transformer· Attention · Medical imaging · Explanation supervision · Image classification

1 Introduction

In recent years, deep neural networks (DNNs) have achieved impressive results on a variety of computer vision tasks, from image classification and object detection to semantic segmentation and image generation. However, their black-box

nature and ever-increasing complexity make their inner workings hard to understand. This sparked a surge of interest in explainable AI (XAI) that provides human-understandable justifications for model behavior [4,9,15,17,22,28,29]. In computer vision, explainability methods typically involve attributing the prediction to relevant parts of the input image, providing insights into the underlying mechanisms and features that contribute to the output. Those usually come in the form of attribution maps derived from the model which can serve as spatial explanatory cues. Extensive research has been devoted to the development of effective methods for extracting indicative attribution maps [2–4, 21, 30, 36]. These methods encompass various approaches such as gradient-based techniques commonly used for Convolutional Neural Networks (CNNs) [5, 26, 31, 32], and attention-based methods in Vision Transformer (ViT) architectures [2, 18].

A major challenge to the development of data-driven algorithms in the medical domain is the limited size of the available datasets [20], largely due to the high cost and complexity of annotating medical data. Combining data from multiple sources, e.g. distinct institutions, imaging equipment, and scanning protocols, may prove challenging due to the resulting presence of distinct imaging features. These features, induced by non-biologic causes, may contribute to overfitting [13,35,37]. Consequently, DNNs trained on medical datasets are susceptible to overfitting, a phenomenon that ultimately undermines their ability to generalize effectively, leading to limited performance in real-world applications.

Since the advent of DNNs, many regularization methods have been proposed to mitigate overfitting [25]. One of the recent approaches is to regularize explanatory cues obtained from the model during training to improve generalization. For image classification models, this strategy may be implemented by guiding attribution maps derived from the model to align with foreground masks of the traget object associated with the class label of the image. By introducing these foreground masks during training, even for a limited subset of the dataset, we enforce consistency between the model's explanatory signals and the location of the class object. This, in turn, "directs" the model's focus to the relevant part of the image, encouraging it to make correct predictions "for the right reasons" [27]. These masks can be acquired through manual spatial annotations, such as segmentation or bounding boxes, or automatically generated by techniques such as self-supervision [12].

While several studies have explored the use of ground truth localization annotations to regularize attribution maps in image classification models, the application of these techniques to ViT-based models remains an area with untapped potential. In this work we introduce a framework termed LGM-ViT (Localization-Guided Medical Vision Transformer), designed for explanation supervision in ViT-based classification models for medical imaging. Following the strategy described above, we devise a framework which utilizes foreground masks during training to regularize attribution maps extracted from ViT. We introduce a novel approach to generate an explanatory attribution map using both the attention matrices and the output embeddings from the final block of the ViT encoder. We construct a loss function comprised of two loss terms: the first term

is a conventional classification loss function, while the second term regularizes attribution maps using foreground masks.

The main contributions of our work are as follows:

- Introducing a general framework for training ViT-based classification models using explanation supervision. While this study focuses on medical imaging datasets, this framework can be applied to other domains.
- Proposing a new method for deriving attribution maps from ViT-based models, tailored for explanation supervision in image classification.
- Conducting comprehensive experiments to validate the effectiveness of the proposed framework. Quantitative results demonstrate the superiority of our approach over existing methods. Qualitative results illustrate the impact of our approach on the model's decision-making process.

2 Related Work

Attribution Methods in computer vision have evolved to provide deeper insights from model decision-making processes [4, 18, 21]. Specifically for ViTs, there are a few methods that are noteworthy. In [1] the attention rollout technique captures attention information from all ViT blocks by extracting attention maps from every ViT encoder block, and condensing the attention heads within each block into a unified map (e.g., through averaging). Subsequently, the maps obtained from all the blocks are combined through multiplication. The Layer-wise Relevance Propagation (LRP) method [24] propagates relevance attributed to the predicted class across the network layers backward to the input image to create a relevancy map. The LRP is extended in the GAE method [10, 11] by extracting the map from each layer based on the attention heads and their gradients.

Explanation Supervision methods have demonstrated their effectiveness in elevating the performance and resilience of Deep Neural Networks (DNNs) across diverse domains and tasks. Specifically for image classification tasks, several studies have explored using ground truth localization annotations to regularize attribution maps in CNN-based classification models [16, 27, 33]. In [27], a method is introduced to efficiently explain and regularize differentiable models by selectively penalizing their input gradients. In [33], saliency maps inferred from the classifier gradients are penalized when these demonstrate poor consistency with lesion segmentation. In [16], an explanation loss is proposed to handle inaccurate boundaries, incomplete regions, as well as the inconsistent distribution of human annotations. Most related to our work is RobustViT proposed in [12] which applies explanation supervision to ViT-based models. The authors introduce an explanation loss that regularizes GAE maps [10] from ViT-based models using masks of the target object to improve their robustness.

3 Method

The proposed LGM-ViT framework aims to boost performance of ViT-based classification models by performing localization supervision using foreground masks of the class object during training. In this view, we propose a novel approach for deriving spatial attribution maps from ViT-based models termed EAFEM (Embedding-Attention Fused Explanation Map), and construct a loss function that promotes consistency between EAFEM maps and their corresponding foreground masks. This strategy guides the model's attention to the relevant parts of the image, promoting accurate predictions based on meaningful features that benefit generalization and model robustness.

3.1 Vision Transformer Architecture

ViT-based models typically consist of three fundamental components: a patch embedding module, a sequence of ViT encoder blocks, and a task-specific head module. Let $I \in \mathbb{R}^{C \times H \times W}$ be the input image to the model, where C is the number of channels (e.g. 3 for RGB images), and H, W are the height and width of the input image, respectively. The input I is divided into non-overlapping patches of size $p \times p$, and fed to the patch embedding module that linearly projects each patch into a 1D token embedding with dimension d. To preserve the positional information of the patches, positional encoding is added to each embedding vector. Following [14], a special classification token, referred to as CLS token, is added. This results in $t = \frac{H}{p} \times \frac{W}{p} + 1$ tokens of dimension d, each representing a patch in the original image, except the CLS token which learns to store class-related information during training. The output of this module is an embedding matrix $E^{(0)} \in \mathbb{R}^{t \times d}$, where the first row is the CLS token, and each remaining row (i.e. token) represents a specific patch originating from the input image.

The embeddings matrix $E^{(0)}$ is fed to a sequence of n ViT encoder blocks. Each block is composed of a multi-head self-attention (MSA) module followed by a multilayer perceptron (MLP). The MSA module functions within a subspace d_h of the embedding dimension d, such that $d_h \cdot h = d$, where h is the number of heads. The self-attention operation of the k^{th} head ($k \in [1, ..., h]$) in the j^{th} block ($j \in [1, ..., n]$) is defined as follows:

$$A_k^{(j)} = \text{softmax}(\frac{Q_k^{(j)} \cdot {K_k^{(j)}}^T}{\sqrt{d_h}}) \tag{1}$$

$$Z_k^{(j)} = A_k^{(j)} \cdot V_k^{(j)} \tag{2}$$

where the $(\cdot)$ operation denotes matrix multiplication. $Q_k^{(j)}, K_k^{(j)}, V_k^{(j)} \in \mathbb{R}^{t \times d_h}$ are sub-spaces of $Q^{(j)}, K^{(j)}, V^{(j)} \in \mathbb{R}^{t \times d}$ (referred to as queries, keys and values) which are three different linear projections of $E^{(j-1)}$, the output embeddings from the previous block. $A_k^{(j)} \in \mathbb{R}^{t \times t}$ is the attention matrix of the k^{th} head in the j^{th} block, representing the pair-wise relations between each two tokens. We

denote $A^{(j)} \in \mathbb{R}^{h\times t\times t}$ as the stacked attention matrices of block j. $Z_k^{(j)} \in \mathbb{R}^{t\times d_h}$ is the output of the self-attention module of the k^{th} head in the j^{th} block. The final output of the MSA module is $Z^{(j)} \in \mathbb{R}^{t\times d}$, the concatenation of $\{Z_1^{(j)}, ..., Z_h^{(j)}\}$. The output $Z^{(j)}$ is then fed to an MLP with one hidden layer. In every ViT block a Layernorm (LN) is applied before each of the two modules (MSA, and MLP), and a residual connection is added after each module. The operations applied in the ViT block can be formulated as follows:

$$Z^{(j)} = \text{MSA}(\text{LN}(E^{(j-1)})) \tag{3}$$

$$Z_{SC}^{(j)} = Z^{(j)} + E^{(j-1)} \tag{4}$$

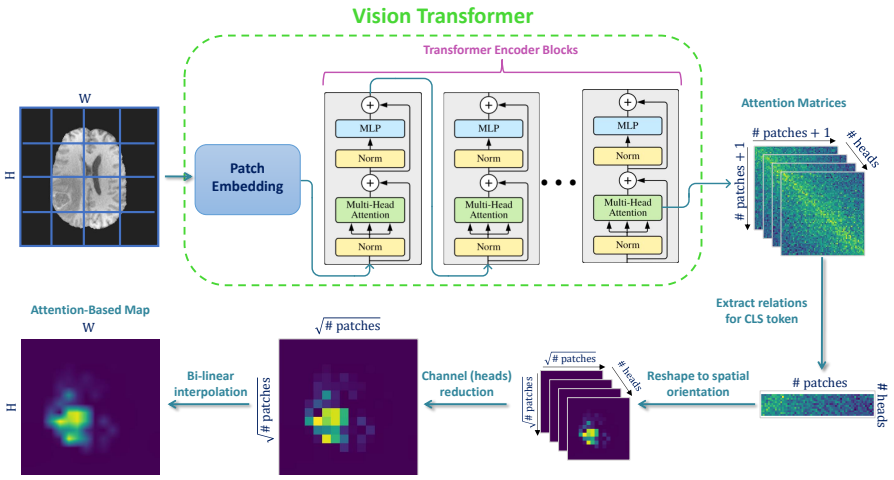
$$E^{(j)} = \text{MLP}(\text{LN}(Z_{SC}^{(j)})) + Z_{SC}^{(j)} \tag{5}$$

where $E^{(j-1)}$ is the output of the previous ViT block, and $Z_{SC}^{(j)}$ stands for the MSA output after adding the skip connection (SC). The structure of the ViT block is illustrated in Fig. 1. The classification head adopted in [14] is an MLP with one hidden layer (not shown in Fig. 1). The input to the MLP head is the CLS embedding token extracted from $E^{(n)}$ (Eq. 5), the output of the final ViT block. For a more detailed explanation of the Transformer architecture, the readers are referred to [14, 34].

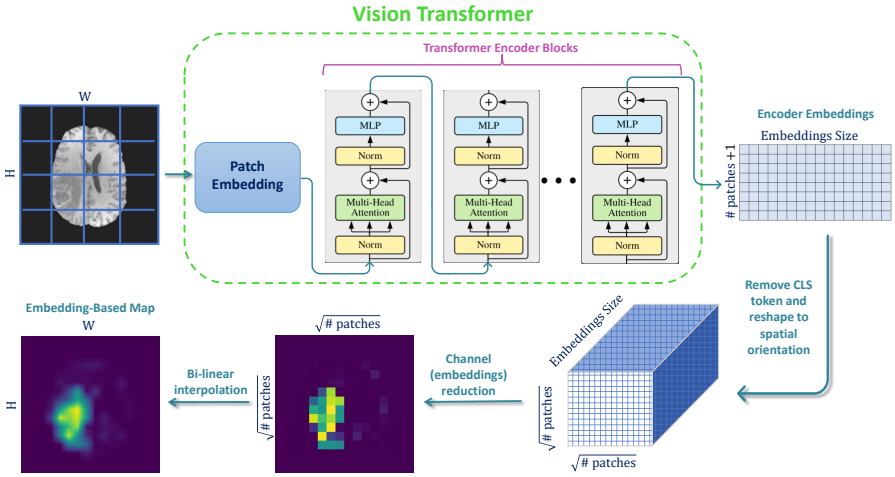
3.2 Attribution Map

The proposed EAFEM method for extracting attribution maps from ViT models utilizes both the attention matrices and the output embeddings from the final ViT block of the model as input sources for the generation of explanatory maps. Each input source is processed separately to generate a spatial attribution map. The attention-based and the embedding-based maps are then fused to obtain the final attribution map referred to as EAFEM. By combining attention information with feature representations, EAFEM offers a comprehensive understanding of how ViT models process visual data and arrive at their decisions. Attention and embedding-based maps are detailed in the following subsections.

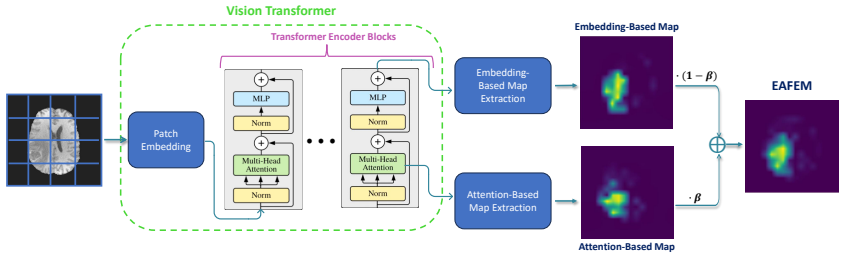
Attention-Based Map. An overview of the attention-based map extraction process is given in Fig. 1a. We extract $A^{(n)} \in \mathbb{R}^{h\times t\times t}$, the attention matrices (Eq. 1) from the final ViT block. Each row in the attention matrix $A_k^{(n)} \in \mathbb{R}^{t\times t}$ corresponds to a specific token, capturing the pairwise connections between that token and all the others. For instance, the i^{th} row represents the relations between the i^{th} token and the other tokens, with the diagonal element (i, i) signifying the relationship of the token with itself. We extract the first row, corresponding to the CLS token, for each of the h attentions matrices, while discarding the diagonal element $(0, 0)$. This results in $A_{CLS}^{(n)} \in \mathbb{R}^{h\times(t-1)}$, describing



(a) Attention-based map extraction process.



(b) Embedding-based map extraction process.



(c) EAFEM extraction process.

Fig. 1: Overview of (a) the attention-based map, (b) the embedding-based map and (c) EAFEM extraction processes.

the relations between the CLS token and the tokens corresponding to the input image patches. By observing from the patch embedding process (Sec. 3.1) that $t-1 = \frac{H}{p} \times \frac{W}{p}$, $A^{(n)}_{CLS}$ can be conveniently reshaped to the spatial orientation of the input image, leading to the multi-head (mh) spatial attention maps, $A^{(n)}_{mh} \in \mathbb{R}^{h \times \frac{H}{p} \times \frac{W}{p}}$. Eventually, a single attention map is obtained by averaging $A^{(n)}_{mh}$ along the heads dimension resulting in $A^{(n)}_{spatial} \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p}}$. The final attention-based map $\mathcal{M}_{attn} \in \mathbb{R}^{H \times W}$, is obtained by bi-linear interpolation of $A^{(n)}_{spatial}$ to the size of the input image.

Embedding-Based Map. An overview of the embedding-based map extraction process is given in Fig. 1b. To generate the embedding-based map, we extract $E^{(n)} \in \mathbb{R}^{t \times d}$, the output of the final ViT block (Eq. 5). We first discard the CLS token from $E^{(n)}$, resulting in $E^{(n)}_{noCLS} \in \mathbb{R}^{(t-1) \times d}$. Then, as previously done for the attention-based map, $E^{(n)}_{noCLS}$ is reshaped to the spatial orientation of the input image, resulting in the embedding-based spatial maps denoted by $E^{(n)}_{spatial} \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times d}$. Similarly to the attention-based map, the final embedding-based map $\mathcal{M}_{embed} \in \mathbb{R}^{H \times W}$ is obtained by averaging along the embedding dimension (d) followed by a bi-linear interpolation to the input image size.

Fusion of the Embedding and Attention Maps. The EAFEM is the weighted mean of the attention-based and embedding-based maps:

$$\mathcal{M}_{fusion} = \beta \mathcal{M}_{attn} + (1-\beta)\mathcal{M}_{embed} \tag{6}$$

where $\mathcal{M}_{fusion}$ is the EAFEM, and $0 \leq \beta \leq 1 \in \mathbb{R}$ is a hyperparameter of the model. An overview of the EAFEM process is depicted in Fig.1c.

3.3 Loss Function

The proposed loss function is composed of two terms. The first term is dedicated to optimizing class prediction. The second term is designed to foster consistency between the attribution maps derived from the model and the foreground masks of the class object. In other words, the first term encourages the model to make correct predictions, improving overall accuracy, while the second term encourages the model to make correct predictions "for the right reasons", enhancing generalization and robustness.

For the first term, we employ the Binary Cross-Entropy (BCE) loss function:

$$\mathcal{L}_{cls} = BCE(\phi(x), y) \tag{7}$$

where x is the input image, $\phi(x) \in [0,1]$ is the prediction of the model, and $y \in \{0,1\}$ is the binary ground truth class.

For the second term, we employ the Kullback–Leibler divergence (KL) loss function:

$$KL(y_{pred}, y_{true}) = Mean\big(y_{true} \odot (\log(y_{true}) - \log(y_{pred}))\big) \tag{8}$$

where the $\odot$ operation denotes the Hadamard product and the $Mean$ operation computes the mean value of the point-wise KL divergence distance map. This function is applied on the attribution map derived from the model and the foreground mask of the class object. 2D Softmax normalization is applied to the attribution map. Following [33], Gaussian smoothing is applied to the binary foreground masks:

$$\mathcal{L}_{loc} = KL(Softmax(\mathcal{M}_{exp}), smth(\mathcal{M}_{fg})) \tag{9}$$

where the $Softmax$ operation is 2D Softmax normalization, the $smth$ operation is Gaussian smoothing, $\mathcal{M}_{exp}$ is the attribution map derived from the model, and $\mathcal{M}_{fg}$ is the foreground mask of the class object. In our framework we use the EAFEM (Eq. 6) introduced in Sec. 3.2 as the attribution map. The localization loss term can be applied to the whole training set or to a subset of it. For training samples having no foreground masks of the class object, or samples in which the class object is absent (e.g. lesion-free slices in medical imaging datasets of lesions), we have $\mathcal{L}_{loc} = 0$. The complete loss function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{loc}\mathcal{L}_{loc} \tag{10}$$

where λ_{loc} is a hyperparameter of the model. An overview of the LGM-ViT loss is illustrated in Fig. 2.

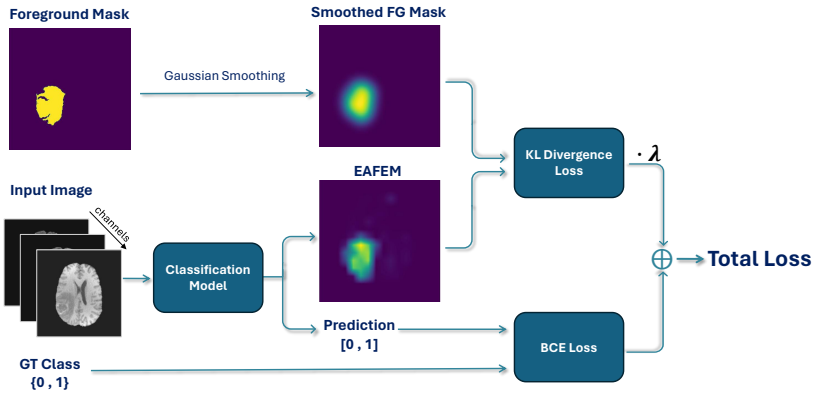


Fig. 2: An overview of the LGM-ViT loss. The input image is fed into the classification model along with the ground truth label and the foreground mask of the class object (if available). The EAFEM and the prediction are extracted from the model. The former is inserted to the localization loss (KL) with a smoothed version of the foreground mask, while the latter is fed to the classification loss (BCE) with the ground truth label. The total loss is the weighted sum of the two loss values, specified by λ.

4 Experiments

We validate the proposed LGM-ViT framework on a binary classification task for two medical imaging datasets for which segmentation ground truth is available. Segmentation annotations are used as foreground masks for the localization supervision (Eq. 9). The vanilla ViT-B/16 [14] is chosen as the baseline model for all the experiments.

4.1 Datasets

BraTS2020. The Brain Tumor Segmentation Challenge 2020 dataset (BraTS2020) [6, 7, 23] is a widely used and comprehensive collection of multimodal magnetic resonance imaging (MRI) brain scans, designed for the evaluation of brain tumor segmentation algorithms. The public training set contains 369 cases, collected from 19 institutions, that were acquired with different protocols, magnetic field strengths, and MRI manufacturers. For each case, four MRI contrasts are provided: a native T1-weighted (T1), a post-contrast T1-weighted (T1ce), a T2-weighted (T2), and a T2 Fluid-Attenuated Inversion Recovery (FLAIR). In this work we use T1, T2, and FLAIR as 3-channel inputs to our models. Preprocessing included co-registration to the T1 modality, skull stripping, and resampling to $1 \times 1 \times 1mm^3$ isotropic resolution, resulting in a common scan size of $155 \times 240 \times 240mm^3$. We refer to the first dimension as slices, and to each slice as distinct model input, such that each input scan comprises 155 slices of dimension 240×240 and the classification model outputs a class prediction for each slice. Annotations provided to each scan consist of four classes: background and healthy tissue (class 0), necrotic and non-enhancing tumor core (NCR/NET - class 1), peritumoral edema (ED - class 2), and enhancing tumor (ET - class 4). Each voxel from the scan is attributed to a single class. In this work we combine classes 1, 2 and 4 into a single non-healthy class. This results in a 2D binary segmentation mask for each slice in the scan, where the 0 class represents background and healthy tissue and the 1 label represents the non-healthy tissue. In addition, since our task is classification per slice, binary ground truth classification labels were assigned to each slice based on its binary segmentation mask (if the sum of the mask was greater than 0 then the class label of the slice is 1, otherwise the class label is 0). For our experiments we randomly split the public training set of this dataset to training, validation and test sets using 70%-10%-20% ratio.

LiTS17. The Liver Tumor Segmentation Challenge 2017 dataset (LiTS17) [8] is a collection of contrast-enhanced 3D abdominal computed tomography (CT) scans used as a benchmark for liver and liver tumor segmentation. The training set contains 131 cases collected from seven clinical sites and acquired with different protocols and manufacturers. The number of slices in each scan ranges from 75 to 987 with a spatial resolution of 512×512 for each slice. The in-plane resolution varies from $0.55mm$ to $1mm$, and the slice spacing between $0.45mm$

Table 1: Binary classification evaluation on the ViT-B/16 [14] model for BraTS2020 and LiTs17 test sets. We compare the performance of our method against the vanilla ViT-B/16 (Baseline), and three competing methods; GradMask [33], RobustViT [12] and RES-G/L [16]. The results present the mean and standard deviation over three runs with different seeds. Best results are marked in bold.

Dataset	Method	F1 Score	Accuracy	AUROC	AP	Cohens Kappa
BraTS2020	Baseline	89.5±0.13	91.2±0.19	96.7±0.20	96.5±0.15	81.9±0.33
	GradMask [33]	89.8±0.24	91.4±0.35	96.7±0.12	96.6±0.05	82.4±0.65
	RobustViT [12]	89.8±0.36	91.3±0.30	96.9±0.04	96.8±0.03	82.2±0.61
	RES-G [16]	90.3±0.58	91.8±0.38	96.9±0.43	96.8±0.39	83.1±0.84
	RES-L [16]	89.6±0.30	91.1±0.23	96.6±0.04	96.6±0.07	81.8±0.48
	LGM-ViT (Ours)	**91.4±0.14**	**92.8±0.14**	**97.3±0.09**	**97.4±0.07**	**85.3±0.26**
LiTS17	Baseline	79.1±0.71	84.7±1.39	93.3±0.51	90.1±1.08	67.0±2.14
	GradMask [33]	81.6±1.81	87.1±1.58	93.7±1.07	90.8±1.44	71.7±3.13
	RobustViT [12]	80.2±0.32	86.6±0.11	93.3±0.34	89.8±0.33	70.0±0.34
	RES-G [16]	82.0±1.58	87.4±1.16	94.0±0.97	90.1±1.56	72.3±2.48
	RES-L [16]	80.3±2.18	85.5±1.58	92.6±1.68	88.1±3.77	68.8±3.34
	LGM-ViT (Ours)	**88.8±0.57**	**92.2±0.56**	**97.2±0.24**	**96.0±0.22**	**82.8±1.07**

and $6mm$. For each scan, liver and liver tumor annotations are provided. In this work we only use the liver segmentation annotations which are in the form of 2D binary masks of the liver for each slice. Following the same methodology applied on the BraTS2020 dataset, binary ground truth classification labels were assigned to each slice based on its binary segmentation mask. Note that in this experiment we focus on the presence/absence of an organ (the liver) in each input slice, and not that of a tumor as in the previous BraTS2020-based experiment. The public training set of this dataset was divided into training, validation and test sets using 70%-10%-20% ratio.

4.2 Implementation Details

Baseline Model. We use the ViT-B/16 [14] as our baseline model for all our experiments. The ViT-B/16 model employs a square patch embedding with patch size of 16, and is composed of 12 sequential ViT encoder blocks with embedding size of 768 and 12 attention heads.

Training. The proposed framework is implemented with PyTorch. Experiments on the BraTS2020 dataset were trained on a single RTX 3090 GPU and experiments on the LiTS17 dataset were trained on a single RTX 5000 GPU. All models were trained from scratch for 25 epochs using ADAM optimizer [19] with an initial learning rate of 0.00001 and a cosine learning rate decay scheduler to a minimum rate of 0.0000001. For both datasets we used a batch size of 32 and resized the input slices to 256×256. The weighting parameter β (Eq. 6) of the EAFEM was optimized using a non-uniform grid search scheme between 0 and 1. The weighting parameter λ_{loc} (Eq. 10) of the loss function was optimized using a non-uniform grid search scheme between 0.01 and 5000. For the BraTS2020

model the β parameter was set to 0.85 and the λ_{loc} to 1000. For the LiTS17 model the β and λ_{loc} were set to 0.95 and 250, respectively.

4.3 Competing Methods

We compared the proposed method to the vanilla ViT-B/16, as well as to its enhanced implementations applying three existing methods used for explanation supervision in image classification: GradMask [33], RobustViT [12], and RES [16]. All the compared methods, applied to ViT-B/16, were trained under the same settings as the LGM-ViT. For comparison fairness, the weighting parameter λ_{loc} was optimized for each of the competing methods using the same scheme employed in the LGM-ViT experiments. In addition, since gradient-based methods are less stable for transformer-based models [3, 12], we replaced the gradient-based maps used for supervision in GradMask [33] and RES [16] with the GAE [10] method employed in RobustViT [12].

4.4 Results

Quantitative Results. Table 1 presents the performance of the proposed LGM-ViT method, alongside the competing approaches, applied to the ViT-B/16 model [14]. We evaluated on two distinct binary classification tasks: (per slice) lesion presence classification using the BraTS2020 dataset and (per slice) liver (organ) presence classification using the LiTS17 dataset. The results represent the average performance over three runs with different seeds. LGM-ViT consistently outperformed all other methods on both datasets across all the evaluation metrics. On the BraTS2020 dataset, our approach achieved an F1 score of 91.4%, accuracy of 92.8%, and an AUROC of 97.3%. This represents an improvement of 1.9%, 1.6%, and 0.6%, respectively, over the baseline ViT model, and an improvement of 1.1%, 1.0%, and 0.4%, respectively, over the next best method (RES-G). LGM-ViT also demonstrated superior performance in terms of Average Precision (97.4%) and Cohen's Kappa (85.3%), surpassing both baseline and competing methods.

The performance gains were even more pronounced on the LiTS17 dataset. LGM-ViT achieved an F1 score of 88.8%, marking a substantial improvement of 9.7% over the baseline ViT and 6.8% over the next best method (RES-G). Similarly, our method attained the highest accuracy (92.2%), AUROC (97.2%), Average Precision (96.0%), and Cohen's Kappa (82.8%) among all compared approaches. The improvements over the baseline ViT were particularly significant, with increases of 7.5%, 3.9%, 5.9%, and 15.8% in accuracy, AUROC, AP, and Cohen's Kappa, respectively. These results demonstrate the effectiveness of LGM-ViT, highlighting its superior ability to enhance the classification capabilities of Vision Transformers in medical imaging applications.

Qualitative Results. In Fig. 3 examples of true positive slices from the BraTS2022 training set are shown. The ground truth annotations (in magenta)

Table 2: Binary classification evaluation for the LGM-ViT with different attribution methods on the LiTS17 test set. The results present the mean and standard deviation over three runs with different seeds. Best results are marked in bold.

Attribution Method	F1 Score	Accuracy	AUROC	AP	Cohens Kappa
None (Baseline)	79.1±0.71	84.7±1.39	93.3±0.51	90.1±1.08	67.0±2.14
Rollout Attention	86.2±0.26	90.3±0.08	96.1±0.21	94.0±0.29	78.7±0.25
GAE	84.3±1.13	88.6±1.34	95.4±0.44	93.2±0.53	75.4±2.37
Attention-based Map	82.8±0.63	87.9±0.63	94.3±0.34	92.3±0.55	73.5±1.18
Embeddings-based Map	87.5±0.71	91.3±0.34	96.3±0.59	95.2±0.67	80.8±0.89
EAFEM	**88.8±0.57**	**92.2±0.56**	**97.2±0.24**	**96.0±0.22**	**82.8±1.07**

and the attention maps of the final block of the ViT are superimposed on top of the input slices. Both the LGM-ViT and the baseline model accurately classified all examples as positive. However, we observe that the correlation between the ground-truth lesion annotations and the attention maps for the LGM-ViT is high, indicating that the LGM-ViT based its prediction on the actual area of the lesion. In contrast, the baseline model's attention maps show weak correlation with the lesion annotations, indicating that it may be relying on other, potentially less relevant, features for classification. This ability of the LGM-ViT to "attend" to the correct anatomical features suggests that it has developed a more robust understanding of the task. By basing its decisions on the most relevant information, the LGM-ViT learns during training to make correct predictions "for the right reasons" enhancing generalization leading to a more robust model.

4.5 Ablation Study

To evaluate the contribution of the EAFEM in LGM-ViT, we conduct an ablation study by replacing it with two leading attribution methods for vision transformers: rollout attention [1] and GAE [10]. Additionally, we replaced the EAFEM with the attention-based and embedding-based maps (Sec. 3.2) separately. We evaluate the LGM-ViT with EAFEM, and the four alternative attribution methods on the LiTS17 dataset. The results are shown in Table 2. The LGM-ViT with the EAFEM outperforms all other attribution methods across all evaluation metrics.

Finally, we assess the impact of the number of scans used for localization supervision during training. Fig. 4 shows the F1 score, accuracy, and AUROC for LGM-ViT as a function of the percentage of training scans used for localization supervision. The results indicate that localization supervision can significantly improve performance even when applied to a limited subset of the training data. Interestingly, while results on the LiTS17 dataset show a gradual increase in performance with more data for localization supervision, on the BraTS2020 dataset

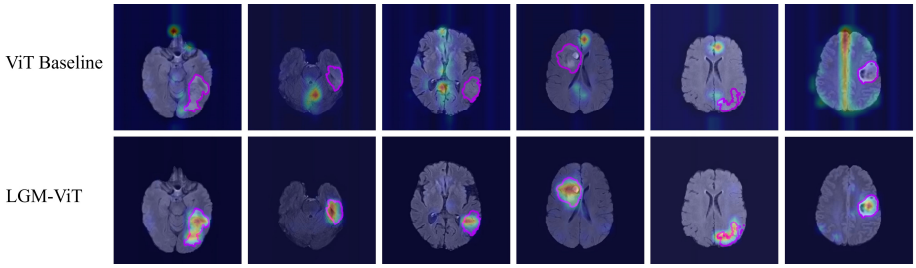


Fig. 3: Examples from the BraTS2022 training set of true positive slices correctly classified by the LGM-ViT and the baseline model. The ground truth annotations (in magenta) and the attention maps of the final block of the ViT are superimposed on top of the input slices.

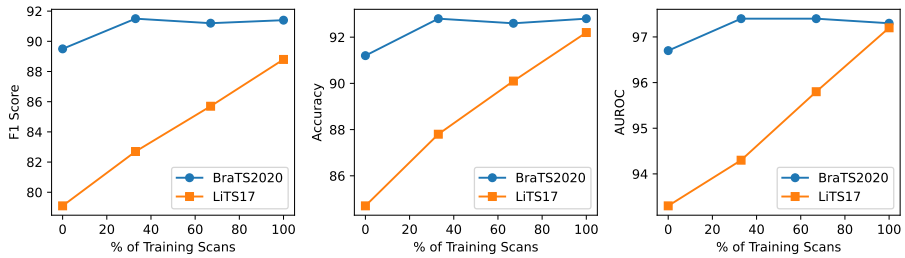


Fig. 4: F1 score, accuracy and AUROC for the LGM-ViT on the BraTS2020 and LiTS17 test sets as a function of the percentage of training scans used for localization supervision during training.

we reach a performance plateau after using only one-third of the data. This suggests that in some cases, localization annotations (used for foreground masks during training) for a small subset of the dataset may be sufficient to maximize the benefits of localization supervision in image classification tasks.

5 Conclusions

In this work we introduced LGM-ViT, a framework designed to enhance the performance and robustness of Vision Transformer models in medical image classification tasks through localization-guided supervision. Our approach integrates a novel method for generating indicative attribution maps with a loss function that promotes consistency between these maps and foreground masks of the class object during training. Experimental results on two challenging medical imaging datasets demonstrate the effectiveness of our approach, underscoring the benefits of localization supervision. LGM-ViT marks a significant advancement in applying Vision Transformers to medical image classification, offering improved performance, interpretability, and robustness. These qualities are essential for developing reliable and trustworthy AI systems in healthcare, where

the stakes are high, and the need for accurate and explainable decision-making is paramount. Although this study is limited to binary classification within the medical domain, our approach is applicable beyond these boundaries. In future research, validation will be extended to multiclass classification, and applied on additional datasets from various domains. Furthermore, the methodology developed in this work can be utilized in applications beyond classification, such as pathology detection, by leveraging the Embedding-Attention Fused Explanation Map (EAFEM) for spatial localization.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020)
2. Achtibat, R., Hatefi, S.M.V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S., Samek, W.: Attnlrp: attention-aware layer-wise relevance propagation for transformers. arXiv preprint arXiv:2402.05602 (2024)
3. Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.R., Wolf, L.: Xai for transformers: Better explanations through conservative propagation. In: International Conference on Machine Learning. pp. 435–451. PMLR (2022)
4. Arras, L., Osman, A., Samek, W.: Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81**, 14–40 (2022)
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
6. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**(1), 1–13 (2017)
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
8. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (LiTS). Medical Image Analysis **84**, 102680 (2023)
9. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021)
10. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 397–406 (2021)
11. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 782–791 (2021)
12. Chefer, H., Schwartz, I., Wolf, L.: Optimizing relevance maps of vision transformers improves robustness. Advances in Neural Information Processing Systems **35**, 33618–33632 (2022)

13. DeGrave, A.J., Janizek, J.D., Lee, S.I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al.: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys **55**(9), 1–33 (2023)
16. Gao, Y., Sun, T.S., Bai, G., Gu, S., Hong, S.R., Liang, Z.: Res: A robust framework for guiding visual explanation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 432–442 (2022)
17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51**(5), 1–42 (2018)
18. Kashefi, R., Barekatain, L., Sabokrou, M., Aghaeipoor, F.: Explainability of vision transformers: A comprehensive review and new perspectives. arXiv preprint arXiv:2311.06786 (2023)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Kiryati, N., Landau, Y.: Dataset growth in medical image analysis research. Journal of Imaging **7**(8), 155 (2021)
21. Komorowski, P., Baniecki, H., Biecek, P.: Towards evaluating explanations of vision transformers for medical imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3726–3732 (2023)
22. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)
23. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
24. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019)
25. Moradi, R., Berangi, R., Minaei, B.: A survey of regularization strategies for deep models. Artificial Intelligence Review **53**(6), 3947–3986 (2020)
26. Nielsen, I.E., Dera, D., Rasool, G., Ramachandran, R.P., Bouaynaya, N.C.: Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. IEEE Signal Processing Magazine **39**(4), 73–84 (2022)
27. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)
28. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**(3), 247–278 (2021)
29. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R.: Explainable AI: interpreting, explaining and visualizing deep learning, vol. 11700. Springer Nature (2019)
30. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q., Nguyen, C.D., Ngo, V.D., Seekins, J., Blankenberg, F.G., Ng, A.Y., et al.: Benchmarking saliency

methods for chest x-ray interpretation. Nature Machine Intelligence **4**(10), 867–878 (2022)
31. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: 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)
32. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
33. Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Gradmask: Reduce overfitting by regularizing saliency. arXiv preprint arXiv:1904.07478 (2019)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
35. Viviano, J.D., Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Saliency is a possible red herring when diagnosing poor generalization. arXiv preprint arXiv:1910.00199 (2019)
36. Watson, M., Hasan, B.A.S., Al Moubayed, N.: Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 875–884 (2022)
37. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Confounding variables can degrade generalization performance of radiological deep learning models. arXiv preprint arXiv:1807.00431 (2018)