A Self-Attention-Guided 3D Deep Residual Network With Big Transfer to Predict Local Failure in Brain Metastasis After Radiotherapy Using Multi-Channel MRI

SEYED ALI JALALIFAR[1], (Member, IEEE), HANY SOLIMAN[2,3,4], ARJUN SAHGAL[2,3,4], AND ALI SADEGHI-NAINI [1,2,3], (Senior Member, IEEE)

[1]Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, ON M3J 1P3, Canada
[2]Physical Sciences Platform, Sunnybrook Research Institute, Sunnybrook Health Sciences Centre, Toronto, ON M4N 3M5, Canada
[3]Department of Radiation Oncology, Odette Cancer Centre, Sunnybrook Health Sciences Centre, Toronto, ON M4N 3M5, Canada
[4]Department of Radiation Oncology, University of Toronto, Toronto, ON M5T 1P5, Canada

CORRESPONDING AUTHOR: A. SADEGHI-NAINI (asn@yorku.ca)

ABSTRACT A noticeable proportion of larger brain metastases (BMs) are not locally controlled after stereotactic radiotherapy, and it may take months before local progression is apparent on standard follow-up imaging. This work proposes and investigates new explainable deep-learning models to predict the radiotherapy outcome for BM. A novel self-attention-guided 3D residual network is introduced for predicting the outcome of local failure (LF) after radiotherapy using the baseline treatment-planning MRI. The 3D self-attention modules facilitate capturing long-range intra/inter slice dependencies which are often overlooked by convolution layers. The proposed model was compared to a vanilla 3D residual network and 3D residual network with CBAM attention in terms of performance in outcome prediction. A training recipe was adapted for the outcome prediction models during pretraining and training the down-stream task based on the recently proposed big transfer principles. A novel 3D visualization module was coupled with the model to demonstrate the impact of various intra/peri-lesion regions on volumetric multi-channel MRI upon the network's prediction. The proposed self-attention-guided 3D residual network outperforms the vanilla residual network and the residual network with CBAM attention in accuracy, F1-score, and AUC. The visualization results show the importance of peri-lesional characteristics on treatment-planning MRI in predicting local outcome after radiotherapy. This study demonstrates the potential of self-attention-guided deep-learning features derived from volumetric MRI in radiotherapy outcome prediction for BM. The insights obtained via the developed visualization module for individual lesions can possibly be applied during radiotherapy planning to decrease the chance of LF.

INDEX TERMS Attention mechanism, brain metastasis, deep learning, stereotactic radiotherapy, therapy outcome prediction.

Clinical and Translational Impact Statement— Early prediction of local failure coupled with visualization of the model's decision basis allows for treatment adjustment on an individual patient basis and facilitates precision oncology for BM patients.

I. INTRODUCTION

About 20% of patients with extracranial malignancies develop brain metastases (BM) [1]. It is estimated that between 70,000 to 400,000 new cases of BM are diagnosed each year in the United States [2]. Because of increasing access to neuroimaging and developments in systemic therapies for patients with metastatic disease, as well as increased physician and patient awareness of BMs, the incidence of BM is expected to increase among cancer patients [1].

The survival of patients with BM depends on timely diagnosis and effective therapy. The major therapeutic options for metastatic brain tumours include surgery, radiation therapy, and/or chemotherapy. Surgical resection is recommended for patient with large solitary tumours in an accessible location [3]. The three principal modalities of radiation therapy (RT) for BM are whole-brain radiation therapy (WBRT), single-fraction stereotactic radiosurgery (SRS), and hypofractionated stereotactic radiotherapy (SRT). While WBRT has been the main treatment for patients with multiple BM [4], there has been a move away from WBRT to SRT and SRS due to adverse side-effects associated with WBRT such as fatigue and cognitive deterioration [5], [6].

Because of various tumour and/or patient-related characteristics such as tumour size, location, and histology as well as the patient's genetics, age and performance status, local response of BM tumours to radiation varies among patients. This is true even when standardized dose/fraction regimens are administrated [7]. The local response to RT is classified as either local control (LC; stable or shrinking tumour that is indicative of a stable disease, partial response, or complete response,) or local failure (LF; enlarging tumour associated with a progressive disease) based on tumour size changes on follow-up structural serial imaging [8]. However, it could take months for a local response to be visible on follow-up scans. Given that the median survival of BM patients following RT can be between 5 months and 4 years [9], [10], early detection of LF after RT potentially permit effective adjustments in treatment that lead to enhanced therapy outcomes, patients' survival, and their quality of life.

Following the successful application of artificial intelligence (AI) methods in diagnostic imaging [11], [12], AI-based cancer imaging analysis is now being used to meet other and more complex clinical challenges [13], [14]. These methodologies have the capacity to uncover previously unknown features from routinely acquired medical images. Quantitative and semi-quantitative features, which are often beyond human perception, can be derived from obtained neuroimaging data. These features can potentially be applied to develop machine learning models to address crucial clinical challenges such as therapy outcome assessment or treatment response prediction. Radiomics is a relatively new transformational research domain that adapts high-throughput approaches for mining of large-scale medical imaging datasets to identify quantitative features (biomarkers) for different diagnostic and prognostic applications [15].

Multiple studies have shown links between radiomic signatures of tumours and their phenotypic, genomic, and proteomic profiles [16]. Several studies have also demonstrated the efficacy of radiomic-based machine learning models in therapy outcome prediction [17], [18], including local response of BM to radiotherapy [19], [20], [21].

Compared to hand-crafted radiomic features, the application of deep learning in medical imaging could possibly address more complicated challenges, particularly when large relevant datasets are available. Deep learning models have shown great promise in recognizing important and distinctive aspects of medical image data in various applications including cancer therapeutics [22], [23], [24]. Deep models, and especially convolutional neural networks (CNNs), can detect complex textural patterns in tissue, distinguish between malignant and benign cells, and possibly derive information from tumour images for therapy outcome prediction [25], [26], [27]. Accordingly, the CNNs can potentially outperform the traditional radiomic models in diagnostic and prognostic applications for precision oncology by detecting patterns in medical images that are not captured by closed-form mathematical definitions of hand-crafted radiomic features [28], [29], [30]. A recent publication from our group shows that the deep-learning features derived from 2D MRI slices outperform the standard clinical variables in predicting radiotherapy outcome in BM [31].

Attention mechanisms in deep learning were introduced in the field of computer vision with the goal of imitating the human visual system's ability to naturally and effectively discover prominent regions in complex scenes [32]. An attention mechanism in a vision system can be thought of as a dynamic selection process that is implemented by adaptively weighing features based on the relevance of the input. Over the past few years, attention mechanism has played an increasingly important role in different computer tasks, including image classification [33], object detection [34], semantic segmentation [35], and 3D vision [36]. The attention mechanism has shown promise in medical imaging analysis, especially when the problem is not as straightforward as generic image classification, where the well-defined object of interest is usually in the image center [37]. A number of previous studies have applied attention mechanisms to provide more powerful architectures capable of catching subtle features covered in medical images. Guan et al. [38] proposed a three-branch attention guided convolution neural network (AG-CNN) which learns from disease-specific regions through a local branch to reduce noise and improve alignment, with a global branch to compensate for the lost discriminative cues by the local branch. Using a fusion branch to combine the local and global cues, their model could achieve a new state-of-the-art performance in classifying images of the ChestX-ray14 dataset [38]. Rao et al. [39] conducted an experimental research to investigate the contribution of various attention mechanisms including squeeze-and-excitation (SE) [33], global context (CG)

[40], and convolutional block attention module (CBAM) [41] to the performance of deep classification models for different imaging modalities including x-ray, MRI, and CT. The experimental results show that the attention mechanisms enable standard CNN models to focus more on semantically important and relevant content within features, with improved area under the receiver operating characteristic (ROC) curve (AUC) for all classification models investigated [39]. Furthermore, the CBAM outperformed the other two attention mechanisms in several experiments on different imaging datasets. Shaik et al. [42] proposed a multi-level attention mechanism for the task of brain tumour classification. The proposed multi-level attention network (MANet) combines spatial and cross-channel attention, focusing on tumour region prioritization while also preserving cross-channel temporal connections found in the Xception backbone's semantic feature sequence [42]. They benchmarked their framework on BraTS [43] and Figshare [44] datasets where their model outperformed several models proposed previously for the brain tumour classification task [42].

This translational study introduces an innovative transformer-convolutional deep learning model for predicting the LC/LF outcome in BM treated with SRT using two-channel MRI acquired at pre-treatment. A novel attention-guided 3D residual network architecture has been developed with embedded self-attention modules [45], [46], [47] and compared with another residual network with 3D CBAM as the attention mechanism. A training recipe has been adapted for the therapy outcome prediction models during pretraining and training the down-stream task based on the recently-proposed big transfer (BiT) principles [48]. A new 3D visualization method has been introduced to illustrate the impact of different regions throughout the lesion volume upon the network's prediction of the therapy outcome. The results demonstrate that incorporating the attention mechanisms into the vanilla 3D residual network improves its performance in outcome prediction considerably, with the self-attention mechanism outperforming the CBAM in terms of accuracy, AUC, and F1-score. Further, the adapted BiT-based recipe for pretraining and hyperparameter tuning improves the deep models' performance in therapy outcome prediction.

II. METHODS AND PROCEDURES
A. DATA ACQUISITION
This study was carried out in compliance with the institutional research ethics board approval from Sunnybrook Health Sciences Centre (SHSC), Toronto, Canada. Data were obtained from 124 BM patients treated with hypo-fractionated SRT (5 fractions). In this study, the baseline treatment-planning MRI including contrast-enhanced T1-weighted (T1w), and T2-weighted-fluid-attenuation-inversion-recovery (T2-FLAIR) images were applied for therapy outcome prediction. The MRI scans were acquired using a 1.5 T Ingenia system (Philips Healthcare, Best, Netherlands) and a 1.5 T Signa HDxt system (GE Healthcare,

Milwaukee, WI, USA). The T1w and T2-FLAIR images had an in-plane image resolution of 0.5 mm and a slice thickness of 1.5 mm and 5 mm, respectively. The treatment-planning tumour contours delineated by expert oncologists as well as the edema contours outlined under their supervision were also included in the dataset. The dataset (124 patients with 156 lesions) was randomly partitioned at patient level into a training set (99 patients with 116 lesions) that was used for model development and optimization, and an unseen test set (25 patients with 40 lesions) that was applied for independent evaluation of the models. From the training set, 10 patients with 15 lesions were randomly selected as the validation set for optimizing the model hyperparameters.

The patients were scanned with MRI after SRT on a two to three-month follow-up schedule. A radiation oncologist and a neuroradiologist determined the local response for each lesion separately after monitoring it on serial MRI using the RANO-BM [8] criteria. The outcome (LC or LF) was determined for each lesion in the last patient follow-up. Serial imaging (including perfusion MRI) and/or histological confirmation were used to diagnose adverse radiation effect (ARE) and distinguish it from progressive disease [49], in accordance with the report by Sneed et al. [50]. Following these criteria, a total of 93 lesions were categorized as LC while 63 lesions were labeled as LF.

B. PREPROCESSING
All MR images were resampled to a size of $512 \times 512 \times 174$ voxels (voxel size: $0.5 \times 0.5 \times 1 \text{ mm}^3$). An affine registration method was used to co-register the T1w and T2-FLAIR images. Skull stripping was performed on all MR images. The voxel intensities in each skull-stripped MRI volume were normalized between 0 and 1. To ensure a lesion-level local outcome prediction the size of the smallest sub-volume enclosing the tumour and edema (lesion) and their 5-mm outer margin [51], was identified for all lesions. A sub-volume of $128 \times 128 \times 83$ voxels was determined as a fit standard to encompass the entire region of interest (ROI) described above for all individual lesions. The standardized sub-volumes were then cropped from the T1w and T2-FLAIR images and concatenated for each lesion as two channels of data, generating the input to the neural networks with a size of $128 \times 128 \times 83 \times 2$ voxels. The ROI masks (tumour + 5-mm margin for T1w; tumour + edema + 5-mm margin for T2-FLAIR) were generated using the tumour and edema contours and applied to mask out the areas outside the ROI for each lesion.

C. NETWORK OVERVIEW
The backbone of the proposed network architecture is a vanilla 3D extension of deep residual networks (ResNets), first introduced by He et al. [52], [53] Instead of learning unreferenced functions, ResNets learn residual functions with reference to the layer inputs. Also, rather than expecting each few stacked layers directly fit a desired underlying mapping, ResNets let these layers fit a residual mapping. Formally,

instead of directly mapping the desired underlying function $H(x)$, the stacked nonlinear layers fit another mapping of $F(x) := H(x) - x$. This way, the original mapping recast into $F(x) + x$. Residual connections allow for increased depth while addressing the vanishing gradient problem and are also easier to optimize [52]. Our vanilla 3D residual network (Figure 1.a) is inspired by the architecture of ResNet-18, but instead of 2D convolution layers, our network employs 3D convolution with kernel size of $7 \times 7 \times 7$ and 3D pooling layers to handle the 3D nature of MRI volume.

To improve the performance of our 3D residual network in processing multi-channel MRI volumes, we explored incorporating two different attention mechanisms into the architecture, CBAM, and self-attention [46]. CBAM is a simple yet effective attention module which infers attention maps along two separate dimensions (channel and spatial) sequentially [41]. The attention maps are then multiplied by the feature tensors to produce the refined feature tensors. Formally, CBAM has two sequential submodules and, given an input feature tensor $x \in R^{X \times Y \times Z \times C}$, it sequentially infers 1D channel attention vector $M_c \epsilon R^C$ and a 3D spatial attention map $M_s \epsilon R^{X \times Y \times Z}$. The developed 3D residual network with CBAM attention is depicted in Figure 1.b. The CBAM attention module (Figure 1.e) is embedded right before the average pooling and fully-connected layers to refine features before classification. The refined features are then flattened and fed to the fully-connected layer for classification.

Additionally, we introduced a novel transformer-convolutional network architecture by incorporating self-attention modules into the 3D residual network (Figure 1.c). The convolution operator in CNNs only conducts local operations and has a local receptive field, but the self-attention mechanism can perform non-local operations and capture long-range dependencies and global information within the input images [54]. The self-attention method is based on the covariance between the elements of feature tensors [55]. Formally, a self-attention function can be described through mapping the input feature tensor to a query, a key, and a value tensor. The tensor mappings are performed using 3D $1 \times 1 \times 1$ convolutions. Each element of the output self-attention feature tensor is a linear weighted sum of the elements of the value tensor. The query tensor defines which "values" to focus on for the learning process, while the key and value tensors carry the transformed features extracted from MRI volume. Given that key is $k(x) = W_k x$, query is $q(x) = W_q x$, and value is $v(x) = W_v x$ where W_k, W_q, W_v are learnable weights of the $1 \times 1 \times 1$ convolution filters and x is the feature tensor from the previous layer, the self-attention map α could be calculated as $\alpha_{i,j} = \frac{\exp(q(x_i)k(x_j)^T)}{\sum_{i=1}^n \exp(q(x_i)k(x_j)^T)}$. Figure 1.f shows the architecture of the 3D self-attention block incorporated into the proposed 3D residual network with self-attention. For performing the matrix multiplications in this block, the query, key, and value tensors ($\in R^{X \times Y \times Z \times C}$) are reshaped into matrices ($\in R^{XYZ \times C}$) and, at the end, reshaped back into tensors of the initial size.

The final $1 \times 1 \times 1$ convolution block ensures that the number of channels of the input and output feature tensors stays the same. The 3D self-attention module facilitates capturing long-range inter/intra slice dependencies, hence is added to the architecture after each residual block to ensure deriving such dependencies along with the convolution layers that mostly capture local features and dependencies. More details on the network architectures have been provided in the Supplementary Materials.

D. BIG TRANSFER AND TRAINING DETAILS

Transfer of pretrained models on the target task improves sample efficiency and simplifies hyperparameter tuning when training deep neural networks [48]. Inspired by the work of Kolesnikov et al. [48], we followed the subsequent scheme for pretraining/training the outcome prediction models:

1. The network was first pretrained on the UCF101 dataset [56] for the task of activity recognition and subsequently on the BraTS dataset [43], [57], [58] for the task of classifying brain tumour types using MRI.

2. During pretraining, all batch normalization [59] layers were replaced with group normalization [60] and weight standardization [61] was used in all convolutional layers. The combination of group normalization and weight standardization with large batches has a significant impact on transfer learning [62]. Also, due to the requirement to update running statistics, batch normalization is detrimental for the transfer [48].

3. During fine-tuning on the main dataset, we used BiT-HyperRule, a heuristic method for hyperparameter selection based on image resolution and number of datapoints as presented by [48]. The models were trained using the stochastic gradient descent (SGD) optimization algorithm with an initial learning rate of 0.003, momentum of 0.9, batch size of 4, and an early stopping based on the validation loss. Data augmentation was performed using horizontal flipping. During fine-tuning, the learning rate was decayed by a factor of 10 at 40%, 60% and 80% of the training steps.

All experiments were performed in Python. The models were developed and evaluated using Keras [63] with Tensor-Flow [64] backend. The performance metrices were calculated using scikit-learn package [65]. the matplotlib library was used [66] for visualization. The models were trained using four GeForce RTX TI 2080 graphic cards. The training process took 5 hours (~33M parameters), 6 hours (~33M parameters) and 10 hours (~42M parameters) for 3D residual network, 3D residual network + CBAM attention, and 3D residual network + self-attention respectively. The total inference time for a single input is 6ms, 7ms, and 12ms for 3D residual network, 3D residual network + CBAM attention, and 3D residual network + self-attention, respectively.

E. VISUALIZATION OF NETWORK DECISION BASIS

A new 3D visualization algorithm was implemented to accompany the outcome prediction framework and show how
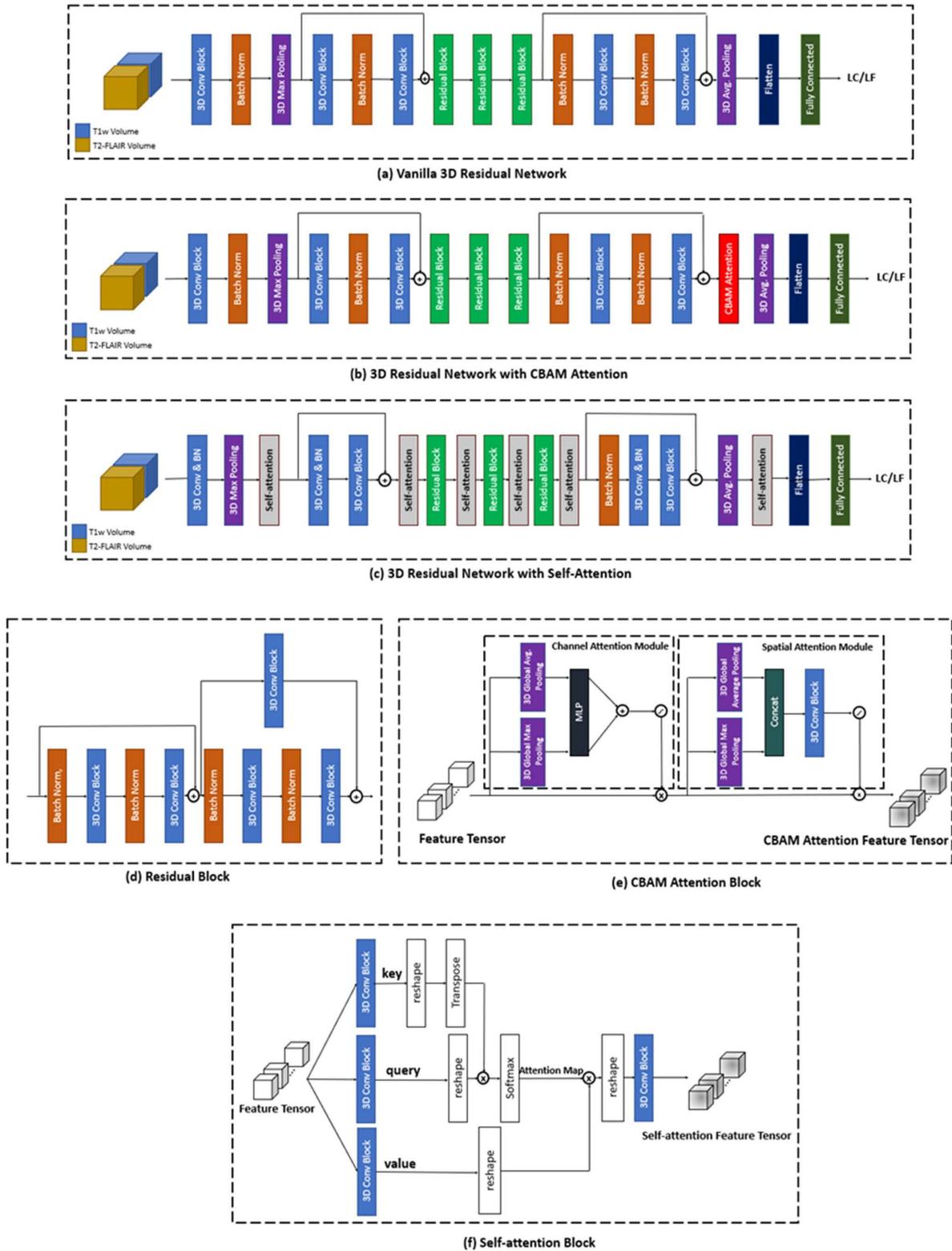
(a) Vanilla 3D Residual Network

(b) 3D Residual Network with CBAM Attention

(c) 3D Residual Network with Self-Attention

(d) Residual Block

(e) CBAM Attention Block

(f) Self-attention Block

FIGURE 1. Architecture of (a) vanilla 3D residual network (baseline), (b) the 3D residual network with CBAM attention, (c) the proposed self-attention-guided 3D residual network, (d) the residual block in 3D residual network, consisting of residual connections, (e) the CBAM attention block consisting of the channel and spatial attention modules, and (f) the self-attention block consisting of the key, query, and value tensors that generates the final self-attention feature tensor.
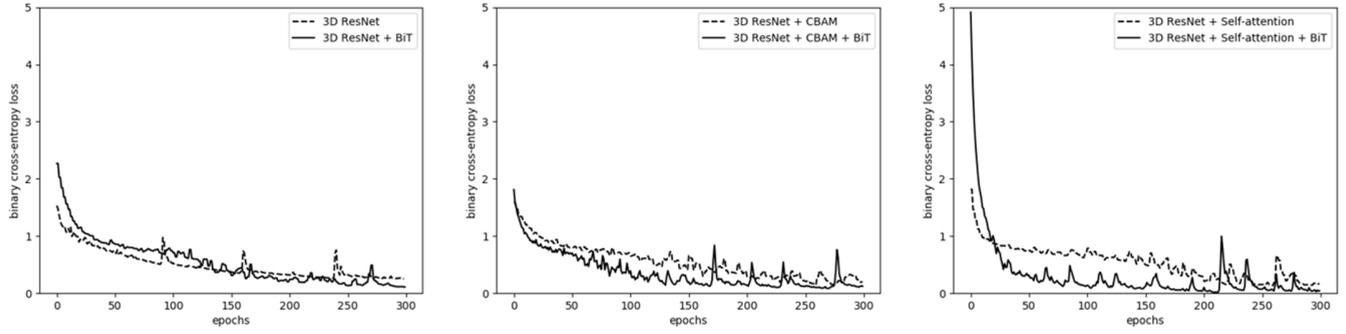
FIGURE 2. Training loss of the three models investigated in this study before and after applying BiT training scheme. Following the BiT training recipe generally led to faster convergence, smaller loss, and better performance overall.

different areas within the volumetric region of interest on the input images contribute to the prediction of network for each lesion. The visualization module provides a 3D heatmap color-coding the relevance of distinct peri-/intra-lesion areas on multi-channel volumetric MRI to the decision of network and may be applied to analyze the reasoning behind the predicted outcome for each case. The applied visualization method combined a modified version of the prediction difference analysis (PDA) with a sliding window analysis approach [67]. A $2 \times 2 \times 1$ voxel sliding black cube ($1 \times 1 \times 1$ mm) was iteratively applied to block a tiny area of the input image. The occluded input was fed to the trained network to predict the associated therapy outcome. In each iteration, the absolute difference in the network's output probability (i.e., $|p_{input} - p_{occluded\_input}|$) was calculated and applied as a measure of contribution of the occluded cube to generate the volumetric heatmap. This method generates a point cloud where each point in the cloud maps to a region within the MRI volume. For 3D visualization of generated heatmap, the heatmap voxels were considered as a point cloud with each point maps to a region within the MRI volume. A surface reconstruction technique was adapted to create a 3D heatmap out of the point cloud on any desired surface within the volumetric ROI. Specifically, the cloud points located on the ROI surface were identified and the normal orientation of the point cloud was calculated at each surface point using a minimum spanning tree with the number of neighbours set to 3 for building the tree [68]. The estimated normal orientations were applied in conjunction with, the Poisson reconstruction technique [69] to build a smooth surface mesh from the point cloud. The Poisson surface reconstruction technique creates a 3D mesh from a dense point cloud by reducing the difference between the surface normal directions of the reconstructed surface and the 3D points in the point cloud [70]. The proposed 3D visualization framework can assist clinicians to get insight into how the network has reached its decision and help to validate the network's decisions by generating meaningful heatmaps. More details on the visualization framework have been provided in Supplementary Materials.

III. RESULTS

The patients (average age: 62 ± 15 years; 40% male and 60% female) had an average tumour size of 2 ± 1.03 cm and an average GPA of 2.2. The demographic and clinical attributes of the patients in this study are presented in Supplementary Table S1.

Figure 2 shows the training loss over 300 epochs for the models in this study before and after applying the BiT training scheme. The vanilla 3D residual network pretrained on the UFC101 and BraTS datasets (without BiT scheme) is the baseline model of this study. In pretraining/training of the models without the BiT recipe, the batch normalization layers were not replaced with group normalization and the weight standardization and BiT-HyperRule were not applied. Table 1 presents the performance of different models investigated in this study for radiotherapy outcome prediction. A careful investigation of Figure 2 and Table 1 demonstrates that incorporating the BiT scheme in development of the deep models for outcome prediction generally improves their performance in terms of convergence, loss, F1-score, and AUC on the independent test set. The F1-score and AUC may be considered the most important metrics presented in Table 1 because of the imbalance exists in the dataset. Specifically, following the BiT training scheme, the models improved their AUC on the test set from 0.83 to 0.84, 0.87 to 0.88, and 0.88 to 0.91 for the 3D residual network, 3D residual network + CBAM attention, and 3D residual network + self-attention, respectively. From a different perspective, incorporating attention mechanisms also improved the model performances in terms of accuracy, AUC, and F1-score. While the vanilla 3D residual network could achieve an F1-score of 75% on the test set, the 3D residual network + CBAM attention improved the F1-score by 2.8%. The F1-score was improved by 3.8% compared to the baseline model by including the self-attention mechanism in the 3D residual network. In particular, the proposed transformer-convolutional network architecture with BiT training demonstrated the best performance in terms of accuracy, AUC, and F1-score, with 8% and 5% improvements in AUC and F1-score, respectively, compared to the baseline model, on

TABLE 1. Results of radiotherapy outcome prediction for different models. Acc: Accuracy; Sens: sensitivity; Spec: specificity.

Network	Validation Set					Independent Test Set				
	Acc.	Sens.	Spec.	AUC	F1-Score	Acc.	Sens.	Spec.	AUC	F1-Score
3D Residual Network	80%	66.7%	88.9%	0.84	72.7%	80%	71%	**87%**	0.83	75%
3D Residual Network + BiT	80%	83.3%	77.8%	0.86	76.9%	80%	82.4%	78.%	0.84	77.8%
3D Residual Network + CBAM Attention	80%	83.3%	77.8%	0.88	76.9%	80%	82.4%	78.2%	0.87	77.8%
3D Residual Network + CBAM Attention + BiT	80%	100%	66.7%	0.88	80%	80%	**88.2%**	73.9%	0.88	78.9%
3D Residual Network + Self-attention	**86.7%**	**83.3%**	**88.9%**	0.89	**83.3%**	**82.5%**	76.5%	**87%**	0.88	78.8%
3D Residual Network + Self-attention + BiT	**86.7%**	**83.3%**	**88.9%**	**0.93**	**83.3%**	**82.5%**	82.4%	82.6%	**0.91**	**80%**
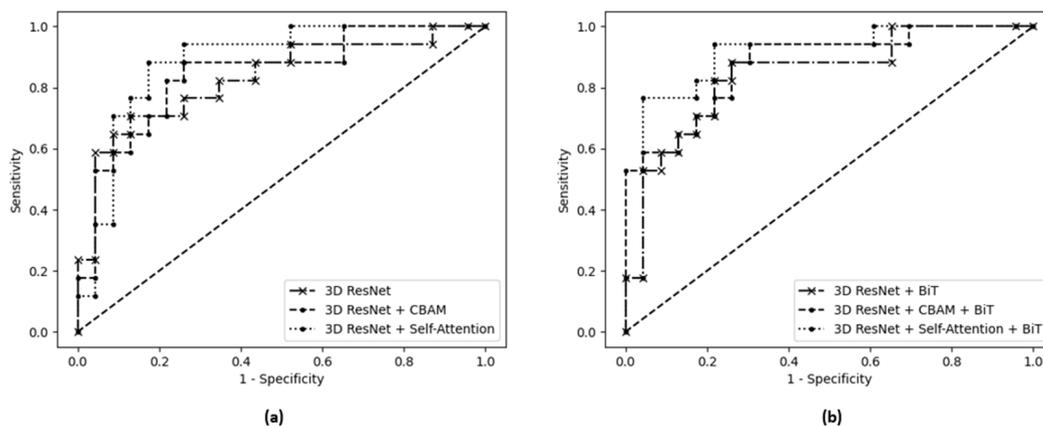
FIGURE 3. The ROC curves for (a) vanilla 3D residual network, 3D residual network + CBAM, and 3D residual network + self-attention, and (b) the same models trained with the BiT scheme.
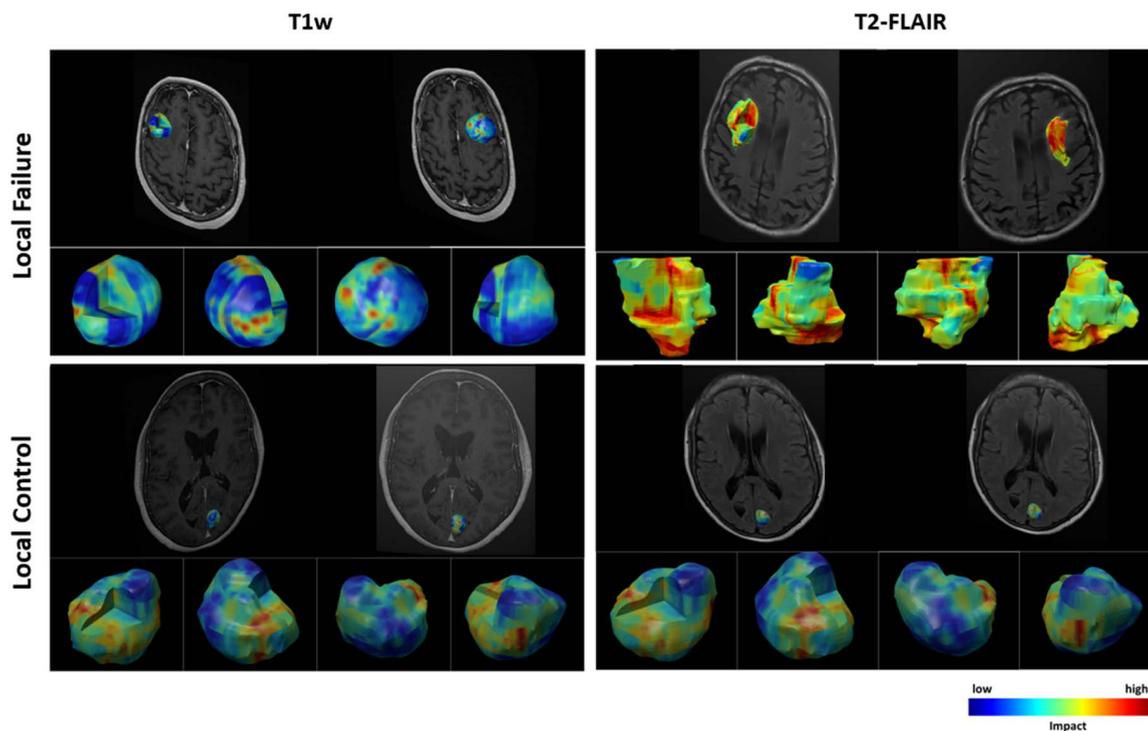

FIGURE 4. 3D visualization heatmaps corresponding to the two input channels (T1w and T2-FLAIR) of the 3D residual network with self-attention and BiT training for two representative lesions, one with an LF (top) and the other one with an LC (bottom) outcome. The user can inspect any desired area on the lesion/margin surface or inside the volumetric ROI and their correspondence with the MRI channels.

the independent test set. This is a considerable improvement in model's performance given the complicated task at hand. Further, the proposed model resulted in the most balanced sensitivity and specificity values compared to the other models despite the imbalanced dataset applied in the study. Figure 3 shows the ROC curves for different models investigated in this study.

An explanation of how an attention mechanism helps improving the performance of the deep models in therapy outcome prediction is as follows: characteristics of different regions within tumour and peritumoral areas on MRI contribute unequally to the likelihood of local response. Several studies demonstrate that tumour margin areas on MRI carry invaluable information regarding the responsiveness of BM to radiotherapy with possibly higher importance for prediction modeling compared to the core areas [51], [71]. Attention mechanisms help the network to capture the subtle information latent within MRI by filtering out irrelevant data and focusing on regions which truly contribute to the network decisions. Moreover, self-attention facilitates capturing long-range dependencies inside the MRI volume, an important concept that simple 3D convolutional layers are not capable of because of their local nature and limited field of view. Comparing the performance of 3D residual network + CBAM attention and the vanilla 3D residual network, the former has outperformed the latter in terms of AUC and F1-Score, although the number of parameters is almost the same for these networks. This shows the benefit of incorporating attention mechanisms in this setting while it may not increase the network complexity considerably. Our further experiments with the 3D residual network with more parameters (∼40M parameters) when extra layers were added to the network resulted in overfitting. This implies that the 3D residual network + self-attention does not simply benefit from the increased number of network parameters but mainly from the structure of the attention layers incorporated. Figure 4 demonstrates the 3D visualization heatmaps for two representative lesions generated using the technique introduced in Section II-E. The heatmaps show the contribution level of different regions within the volumetric ROI on the prediction of the proposed attention-guided model for each lesion in terms of local outcome. The 3D heatmaps can aid clinicians to examine the lesion volume thoroughly and inspect impactful regions for a predicted outcome which can eventually support their decision making in assessment, diagnosis, and treatment planning.

IV. DISCUSSION AND CONCLUSION

An end-to-end 3D convolutional deep learning architecture with self-attention was introduced in this study to predict the local outcome in BM after radiotherapy. By employing 3D residual blocks in the proposed model, we investigated the possibility of early prediction of LF in BM treated with SRT using T1w and T2-FLAIR MRI volumes acquired at baseline. We further investigated the effect of incorporating attention mechanisms into the 3D residual network. The results show

that the proposed model with self-attention mechanism outperforms the vanilla 3D residual network and the 3D residual network with CBAM attention in terms of accuracy, AUC, and F1-score. The proposed architecture combines residual learning with self-attention mechanism, allowing for full utilization of both global and local information while avoiding information loss. Specifically, the self-attention mechanism in the model takes into account long-range dependencies in the input MRI volumes while the residual connections allow the extracted information to persist throughout the network. We further improved the model's performance by following the BiT scheme for pretraining and hyperparameter tuning. A 3D visualization module was developed and coupled with the framework to show the important areas of lesion on MRI with higher impact on the model's decision. The visualization results confirm the findings of previous studies that the characteristics of tumour/lesion margin areas on T1w and T2-FLAIR images are important for predicting local outcome in BM treated with radiation therapy. In particular, these regions are among the high-impact regions to the predictions made by the proposed deep learning model with more attention gained from the model for therapy outcome prediction.

The findings of this study demonstrate the feasibility of early prediction of radiotherapy outcome for BM using only the features extracted from multi-modal MRI volumes. This study highlights the effect of adding attention mechanism to deep networks and the importance of pretraining in transferring knowledge to the fine-tuning step. When dealing with large models and large datasets (which is usually the case during pretraining) adhering to the BiT recipe allows for optimized training during the up-stream task and a computationally inexpensive fine-tuning protocol during the down-stream task to avoid a complex and costly hyper-parameter search. The obtained results are promising and encourage future studies on larger patient populations. The results of this study were obtained using an independent test set that was kept unseen during the model training and optimization. However, for a more rigorous evaluation of the efficacy and robustness of the models in the clinic, further investigations should be performed on larger patient cohorts and preferably with multi-institutional data.

REFERENCES

[1] P. Sacks and M. Rahman, "Epidemiology of brain metastases," *Neurosurg. Clinics North Amer.*, vol. 31, no. 4, pp. 481–488, Oct. 2020, doi: 10.1016/j.nec.2020.06.001.

[2] N. Lamba, P. Y. Wen, and A. A. Aizer, "Epidemiology of brain metastases and leptomeningeal disease," *Neuro-Oncol.*, vol. 23, no. 9, pp. 1447–1456, Sep. 2021, doi: 10.1093/neuonc/noab101.

[3] C. M. Carapella, N. Gorgoglione, and P. A. Oppido, "The role of surgical resection in patients with brain metastases," *Current Opinion Oncol.*, vol. 30, no. 6, pp. 390–395, Nov. 2018, doi: 10.1097/CCO.0000000000000484.

[4] P. D. Brown, M. S. Ahluwalia, O. H. Khan, A. L. Asher, J. S. Wefel, and V. Gondi, "Whole-brain radiotherapy for brain metastases: Evolution or revolution?" *J. Clin. Oncol.*, vol. 36, no. 5, pp. 483–491, Feb. 2018, doi: 10.1200/JCO.2017.75.9589.

[5] P. D. Brown et al., "Effect of radiosurgery alone vs radiosurgery with whole brain radiation therapy on cognitive function in patients with 1 to 3 brain metastases: A randomized clinical trial," *J. Amer. Med. Association*, vol. 316, no. 4, p. 401, Jul. 2016, doi: 10.1001/jama.2016.9839.

[6] P. D. Brown et al., "Postoperative stereotactic radiosurgery compared with whole brain radiotherapy for resected metastatic brain disease (NCCTG N107C/CEC·3): A multicentre, randomised, controlled, phase 3 trial," *Lancet Oncol.*, vol. 18, no. 8, pp. 1049–1060, Aug. 2017, doi: 10.1016/S1470-2045(17)30441-2.

[7] F. R. Vogenberg, C. I. Barash, and M. Pursel, "Personalized medicine: Part 1: Evolution and development into theranostics," *Pharmacol. Ther.*, vol. 35, no. 10, pp. 560–576, Oct. 2010.

[8] N. U. Lin et al., "Response assessment criteria for brain metastases: Proposal from the RANO group," *Lancet Oncol.*, vol. 16, no. 6, pp. e270–e278, Jun. 2015, doi: 10.1016/S1470-2045(15)70057-4.

[9] P. W. Sperduto et al., "Estimating survival in patients with lung cancer and brain metastases," *JAMA Oncol.*, vol. 3, no. 6, p. 827, Jun. 2017, doi: 10.1001/jamaoncol.2016.3834.

[10] P. W. Sperduto et al., "Estimating survival in melanoma patients with brain metastases: An update of the graded prognostic assessment for melanoma using molecular markers (melanoma-molGPA)," *Int. J. Radiat. Oncol.*, vol. 99, no. 4, pp. 812–816, Nov. 2017, doi: 10.1016/j.ijrobp.2017.06.2454.

[11] A. S. Panayides et al., "AI in medical imaging informatics: Current challenges and future directions," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1837–1857, Jul. 2020, doi: 10.1109/JBHI.2020.2991043.

[12] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[13] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, "Integrated multi-omics analyses in oncology: A review of machine learning methods and tools," *Frontiers Oncol.*, vol. 10, Jun. 2020, doi: 10.3389/fonc.2020.01030.

[14] A. Tiwari, S. Srivastava, and M. Pant, "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019," *Pattern Recognit. Lett.*, vol. 131, pp. 244–260, Mar. 2020, doi: 10.1016/j.patrec.2019.11.020.

[15] J. P. B. O'Connor et al., "Imaging biomarker roadmap for cancer studies," *Nature Rev. Clin. Oncol.*, vol. 14, no. 3, pp. 169–186, Mar. 2017, doi: 10.1038/nrclinonc.2016.162.

[16] H. J. W. L. Aerts et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, no. 1, p. 4006, Sep. 2014, doi: 10.1038/ncomms5006.

[17] S. Rizzo et al., "Radiomics: The facts and the challenges of image analysis," *Eur. Radiol. Exp.*, vol. 2, no. 1, pp. 1–8, 2018, doi: 10.1186/s41747-018-0068-z.

[18] S. Bae et al., "Radiomic MRI phenotyping of glioblastoma: Improving survival prediction," *Radiol.*, vol. 289, no. 3, pp. 797–806, Dec. 2018, doi: 10.1148/radiol.2018180200.

[19] E. Karami, M. Ruschin, H. Soliman, A. Sahgal, G. J. Stanisz, and A. Sadeghi-Naini, "An MR radiomics framework for predicting the outcome of stereotactic radiation therapy in brain metastasis," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 1022–1025, doi: 10.1109/EMBC.2019.8856558.

[20] C.-Y. Liao et al., "Enhancement of radiosurgical treatment outcome prediction using MRI radiomics in patients with non-small cell lung cancer brain metastases," *Cancers*, vol. 13, no. 16, p. 4030, Aug. 2021, doi: 10.3390/cancers13164030.

[21] A. Mouraviev et al., "Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery," *Neuro. Oncol.*, vol. 22, pp. 797–805, Jan. 2020, doi: 10.1093/neuonc/noaa007.

[22] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, Aug. 2021, doi: 10.1016/j.neucom.2021.03.035.

[23] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: A survey," *Evol. Intell.*, vol. 15, no. 1, pp. 1–22, Mar. 2022, doi: 10.1007/s12065-020-00540-3.

[24] M. Wang, Q. Zhang, S. Lam, J. Cai, and R. Yang, "A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning," *Frontiers Oncol.*, vol. 10, p. 2177, Oct. 2020, doi: 10.3389/fonc.2020.580919.

[25] E. Wetzer, J. Gay, H. Harlin, J. Lindblad, and N. Sladoje, "When texture matters: Texture-focused CNNs outperform general data augmentation and pretraining in oral cancer detection," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 517–521, doi: 10.1109/ISBI45749.2020.9098424.

[26] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, "From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 132–160, Jul. 2019, doi: 10.1109/MSP.2019.2900993.

[27] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jul. 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[28] A. Diamant, A. Chatterjee, M. Vallières, G. Shenouda, and J. Seuntjens, "Deep learning in head & neck cancer outcome prediction," *Sci. Rep.*, vol. 9, no. 1, p. 2764, Dec. 2019, doi: 10.1038/s41598-019-39206-1.

[29] S. J. Cho, L. Sunwoo, S. H. Baik, Y. J. Bae, B. S. Choi, and J. H. Kim, "Brain metastasis detection using machine learning: A systematic review and meta-analysis," *Neuro-Oncol.*, vol. 23, no. 2, pp. 214–225, Feb. 2021, doi: 10.1093/neuonc/noaa232.

[30] Y. J. Cha et al., "Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks," *Anticancer Res.*, vol. 38, no. 9, pp. 5437–5445, Sep. 2018, doi: 10.21873/anticanres.12875.

[31] S. A. Jalalifar, H. Soliman, A. Sahgal, and A. Sadeghi-Naini, "Predicting the outcome of radiotherapy in brain metastasis by integrating the clinical and MRI-based deep learning features," *Med. Phys.*, pp. 1–12, Jul. 2022, doi: 10.1002/mp.15814.

[32] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, pp. 331–368, Sep. 2022, doi: 10.1007/s41095-022-0271-y.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[35] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149, doi: 10.1109/CVPR.2019.00326.

[36] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, Jun. 2021, doi: 10.1007/s41095-021-0229-5.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[38] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*.

[39] A. Rao, J. Park, S. Woo, J.-Y. Lee, and O. Aalami, "Studying the effects of self-attention for medical image analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3409–3418, doi: 10.1109/ICCVW54120.2021.00381.

[40] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980, doi: 10.1109/ICCVW.2019.00246.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[42] N. S. Shaik and T. K. Cherukuri, "Multi-level attention network: Application to brain tumor classification," *Signal, Image Video Process.*, vol. 16, no. 3, pp. 817–824, Apr. 2022, doi: 10.1007/s11760-021-02022-0.

[43] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," Nov. 2018, *arxiv:1811.02629*.

[44] J. Cheng, "Brain tumor dataset," Dataset, 2017, doi: 10.6084/m9.figshare.1512427.v5.

[45] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11, [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053 c1c4a845aa-Paper.pdf

[47] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 7354–7363, [Online]. Available: https://proceedings.mlr.press/v97/zhang19d.html

[48] A. Kolesnikov et al., "Big transfer (BiT): General visual representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 491–507.

[49] M. T. Truong et al., "Results of surgical resection for progression of brain metastases previously treated by gamma knife radiosurgery," *Neurosurgery*, vol. 59, no. 1, pp. 86–97, Jul. 2006, doi: 10.1227/01.NEU.0000219858.80351.38.

[50] P. K. Sneed et al., "Adverse radiation effect after stereotactic radiosurgery for brain metastases: Incidence, time course, and risk factors," *J. Neurosurg.*, vol. 123, no. 2, pp. 373–386, Aug. 2015, doi: 10.3171/2014.10.JNS141610.

[51] E. Karami et al., "Quantitative MRI biomarkers of stereotactic radiotherapy outcome in brain metastasis," *Sci. Rep.*, vol. 9, no. 1, p. 19830, Dec. 2019, doi: 10.1038/s41598-019-56185-5.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[54] X. Yang, "An overview of the attention mechanisms in computer vision," *J. Phys., Conf.*, vol. 1693, no. 1, Dec. 2020, Art. no. 012173, doi: 10.1088/1742-6596/1693/1/012173.

[55] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[56] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," Center Res. Comput. Vis., Univ. Central Florida, Orlando, FL, USA, Tech. Rep. CRCV-TR-12-01, Nov. 2012, doi: 10.48550/arXiv.1212.0402.

[57] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[58] S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, Dec. 2017, Art. no. 170117, doi: 10.1038/sdata.2017.117.

[59] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[60] Y. Wu and K. He, "Group normalization," *Int. J. Comput. Vis.*, vol. 128, pp. 742–755, Mar. 2020, doi: 10.1007/s11263-019-01198-w.

[61] S. Qiao, H. Wang, C. Liu, W. Shen, and A. Yuille, "Micro-batch training with batch-channel normalization and weight standardization," 2019, *arXiv:1903.10520*.

[62] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[63] F. Chollet. (2015). *Keras.* GitHub. [Online]. Available: https://github.com/fchollet/keras

[64] M. Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/

[65] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[66] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[67] L. M Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," 2017, *arXiv:1702.04595*.

[68] G. M and R. R, "An improved method for segmentation of point cloud using minimum spanning tree," in *Proc. Int. Conf. Commun. Signal Process.*, Apr. 2014, pp. 833–837, doi: 10.1109/ICCSP.2014.6949960.

[69] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. 4th Eurographics Symp. Geometry Process.*, vol. 7, 2006, pp. 1–10.

[70] A. Maiti and D. Chakravarty, "Performance analysis of different surface reconstruction algorithms for 3D reconstruction of outdoor objects from their digital images," *SpringerPlus*, vol. 5, no. 1, p. 932, Dec. 2016, doi: 10.1186/s40064-016-2425-9.

[71] D. A. Hardesty and P. Nakaji, "The current and future treatment of brain metastases," *Frontiers Surg.*, vol. 3, p. 30, May 2016, doi: 10.3389/fsurg.2016.00030.

● ● ●