





A Siamese Vision Transfer Architecture for Prediction of Brain Tumour Responses during Therapy

Xiaohong W. Gao¹ (✉) , Chia-Hui Chien¹, Guan-Lin Liu¹, and Jyh-Cheng Chen² 

¹ Department of Computer Science, Middlesex University, Hendon, London NW4 4BT, UK
x.gao@mdx.ac.uk

² Department of Biomedical Imaging and Radiological Sciences, National Yang Ming Chiao
Tung University, Taipei, Taiwan

Abstract. This paper presents the results of prediction of therapeutic responses for patients who undergo chemo or radio therapy for the treatment of brain tumour. This work is in response to Task 11 of 2025 BraTS Brain Tumor Progression Challenge organized in conjunction with MICCAI 2025 (<https://conferences.miccai.org/2025/en/>). In this competition, a Siamese Vision Transformer (SViT) is applied, which allows the inferences between baseline state, i.e. after tumour resection or initial treatment and later tumour development. The backbone model is CMT-Ti. Inspired by the work of MuSiC_ViT for x-ray chest disease detection, this SViT system accomplishes four classification of therapeutic responses, which are Complete Response (CR), Partial Response (PR), Stable Disease (SD) and Progressive Disease (PD). Overall, based on the available training dataset with 91 patients, 90% accuracy can be achieved. For the test dataset, the model SViT has achieved top 2 performance for Task 11.

Keywords: Siamese Neural network · Vision Transformers · BraTS challenges · Brain Tumour Therapy

1 Introduction

1.1 A Subsection Sample

This 2025 BraTS Challenge is the continuation of previous competitions [1–4] and has attracted significant numbers of participants. Technically, the Task-11 of the Challenge in essence is a classification problem to identify the status of response after a patient has undergone a therapy. The challenge here is that this response is related to the baseline tumour status, i.e. after initial tumour resection or therapy. Hence, a Siamese convolutional neural network (CNN) appears to be appropriate, which is designed to compare the similarity between two inputs. Within this architecture, shared weights are used, implying the same neural network processes both inputs in parallel, learning comparable representations. In addition, a Vision Transformer (ViT) refers to a deep learning architecture that adapts the transformer model, originally designed for natural language

processing, to image data. Instead of using convolutional layers like CNNs, ViTs treat images as sequences of patches, enabling them to capture long-range dependencies and global context more effectively. When CNN meets ViT (CMT-Tiny), a lightweight network is created to take advantage of both CNN and ViT, transforming into a compact, efficient model for image classification and related tasks [5].

2 Methodology

2.1 Deep Learning Network

Inspired by the work by Cho et al. [6], a Siamese Vision Transformer (SViT) network is employed for this task as presented with CMT-Ti as backbone model in Fig. 1, the SViT architecture is illustrated and embeds two networks that share the same weights. The loss functions are calculated based on three metrics, attention mechanism that is applied to ViT networks, visual similarity between baseline and follow-ups and anatomic region similarity check disease status.

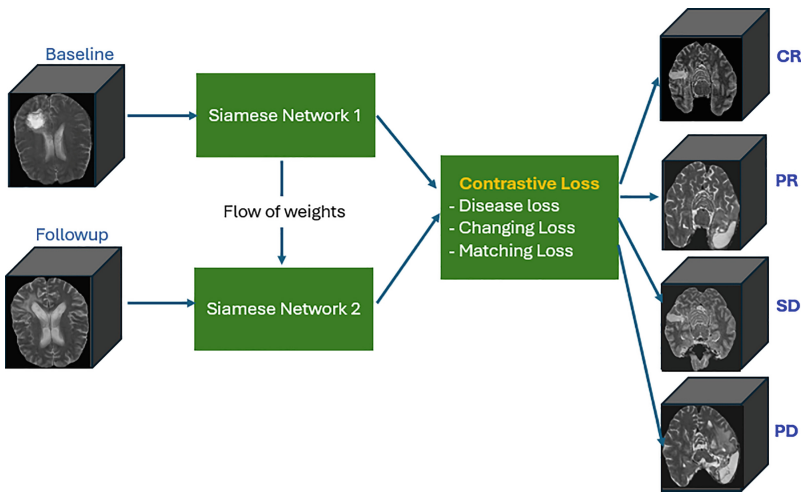


Fig. 1. The architecture of Siamese Vision Transformer (SViT) applied in this study.

In Fig. 1, the SviT network processes a pair of baseline and follow-up images. The forward model employs the architecture of CMTs [5] as an encoder. Dissimilar to a CNN-based model, a vision transformer (ViT) model uses a large reception field [7] thanks to the application of an attention mechanism. While a ViT is mainly trained for capturing global image features, multi-scale features can be obtained via CMT [5] to ensure low-resolution features are captured, which is particularly important for medical images where sub-changes are crucial for accurate diagnosis. Towards this end, CMT utilises depth-wise convolution and multi-head self-attention to efficiently capture local and global structure information through the integration of CNN and ViT.

Similar to the work conducted in [6], Fig. 1 performs a multi-task learning. One task is to classify a pair of baseline and follow-up 2D images into are no-response/complete-response (CR), no-response/partial-response (PR), normal/stable-disease (SD) and normal/progressive-disease (PD). Another is to compare anatomic region similarities. Two cross-entropy loss functions are calculated to distinguish normal and response classes (i.e., CR, PR, SD and PD labels) from baseline (y_b) and follow-up (y_{fu}) MR brain images.

The overall loss is formulated in Eq. (1) with three loss functions to assess response/no-response classes for each patient, by which the performance is enhanced when factors λ_1 , λ_2 , and λ_3 were set to 1, 0.1, and 0.01 respectively empirically.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{response} + \lambda_2 \mathcal{L}_{disease} + \lambda_3 \mathcal{L}_{matching} \quad (1)$$

In Eq. (1), the $\mathcal{L}_{response}$ is calculated in Eq. (2).

$$\mathcal{L}_{response} = CE\left(y_{response}, S\left(W_3 f(x_b) \oplus f(x_{fu})\right)\right) \quad (2)$$

$$\mathcal{L}_{disease} = CE(y_b, S(W_1 f(x_b))) + CE(y_{fu}, S(W_2 f(x_{fu}))) \quad (3)$$

In Eqs. (2) and (3), $y_{response}$ indicates the label for the response/no-response classes of image pairs, and \oplus denotes vector-wise concatenation. The weights of W_1 , W_2 , and W_3 denote fully connected layers.

Where $S(\cdot)$ is the softmax function and is computed using Eq. (4).

$$S(x_i) = \frac{e^{x_i}}{\sum_{k=1}^K e^{x_k}}, \text{ for } i = 1, \dots, K. \quad (4)$$

K refers to the number of data samples.

The cross-entropy (CE) loss function is used to determine the four classes of an MR image pair and is computed as Eq. (5).

$$CE(y, f(x)) = \sum_i y_i \log f(x_i) \quad (5)$$

Similar to the work by Cho et al. [5], the matching loss function denotes to the anatomy-matching module (AMM) that matches the associated features maps from two brain images as presented in Fig. 2. The AMM comprises a feature extraction part (FEP) and a channel recalibration part (CRP). The loss function is calculated in Eq. (6).

$$\mathcal{L}_{matching} = 2 - 2 \frac{1}{n} \sum_{i=1}^n (K_i \bullet Q_i) / K_i Q_i \quad (6)$$

Where K_i and Q_i are the features generated in ViT as key and query focusing on the similar regions through similarity modeling of the K and Q created by the FEP in the two paired images, representing cosine similarity formula. In Eq. (5), n is 4 as shown in Fig. 2 below.

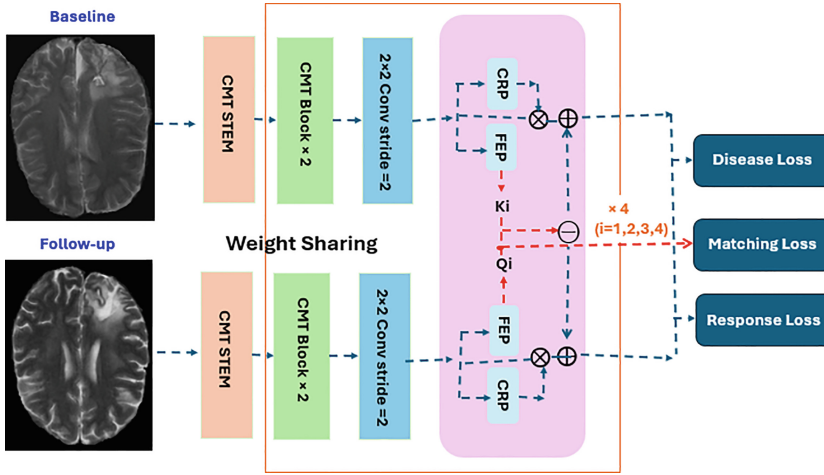


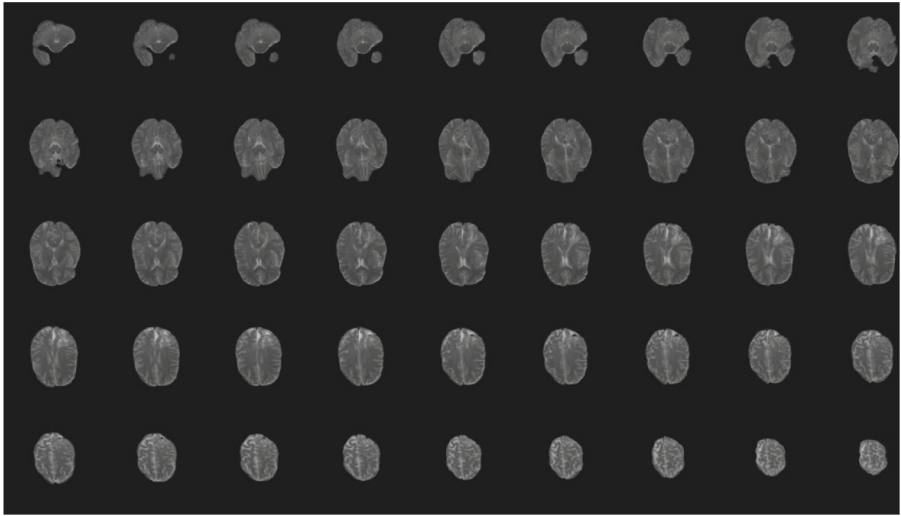
Fig. 2. The workflow of calculation of loss functions for SviT network.

2.2 Datasets

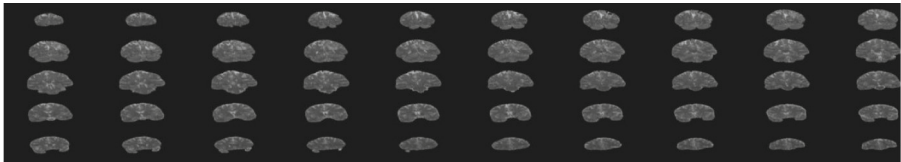
Data are collected from the BraTS 2025 Challenge [1] and are in 3D form of MR images. Due to the variations of volume sizes, e.g. many images have one direction containing only 24 slices, the training and testing take place based on 2D frames. In this study, only registered MR dataset is employed. At each z direction, 3 frames are selected containing the top 3 largest lesions, calculated according to the segmentation masks. For testing with segmentation information, this is estimated based on the gray level intensity. Figure 3 demonstrates a sample data set for a patient from 3 direction views, Axial (top), Coronal(middle) and Sagittal (bottom), where the circled 3 frames in each direction are applied for training. If the width or height is less than 100 pixels (1 pixel = 500 μm), then this frame is discarded.

Figure 4 demonstrates the selection of top 3 frames in each direction to be applied for training. For z direction, if the frame numbers are less than 100, the top 3 will be the first 3 frames with the largest sum of pixel values. If the frame numbers are larger than 100, the 3 consecutive frames will bear little differences. Hence the incremental of selection will be the whole number of $(Z + 100)/100$. For example, if $Z = 512$, then the incremental will be 6, i.e. the top 3 will be the slices top 1, 7, and 13. In Fig. 3, the segmentation marks are super-imposed with red referring to ‘Tumor’ and orange ‘other lesions (e.g. swelling)’ as provided by the dataset.

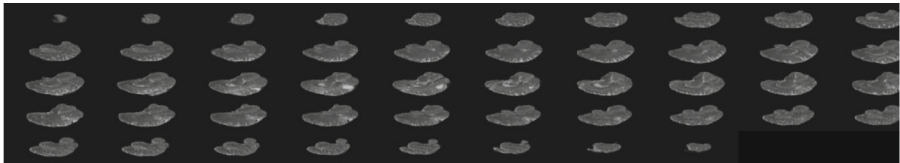
Figure 5 illustrates the paired samples for training with baseline (SviT Network 1 in Fig. 1) and follow-up images (SviT Network 2 in Fig. 1).



Axial



Coronal



Sagittal

Fig. 3. Three directional view of a 3D T2 MR image. Top: Axial; Middles: Coronal; bottom: Sagittal.

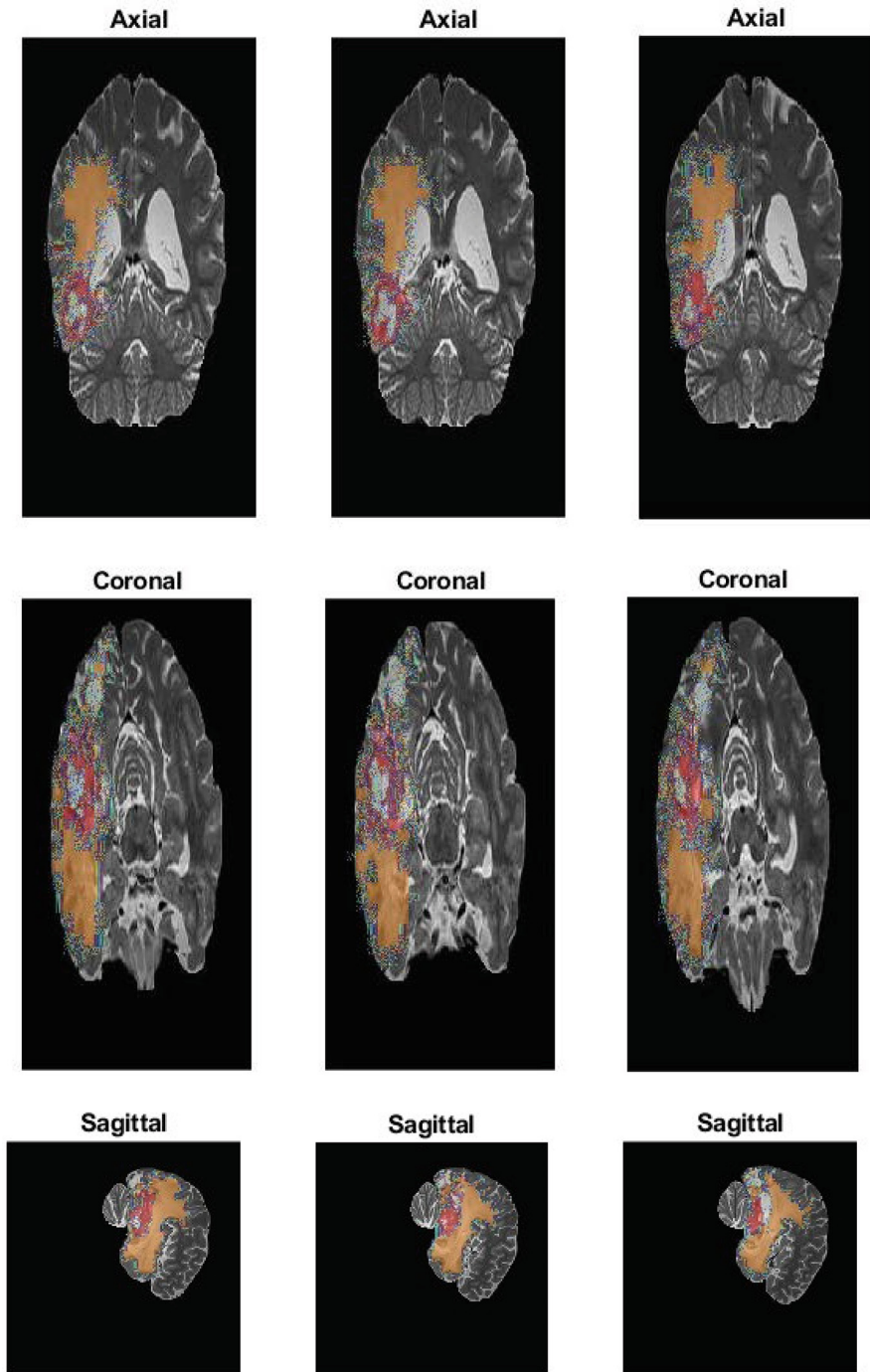


Fig. 4. The selected frames for training and testing from the patient data shown in Fig. 2. The red colour refers to tumour whereas orange the other lesions.

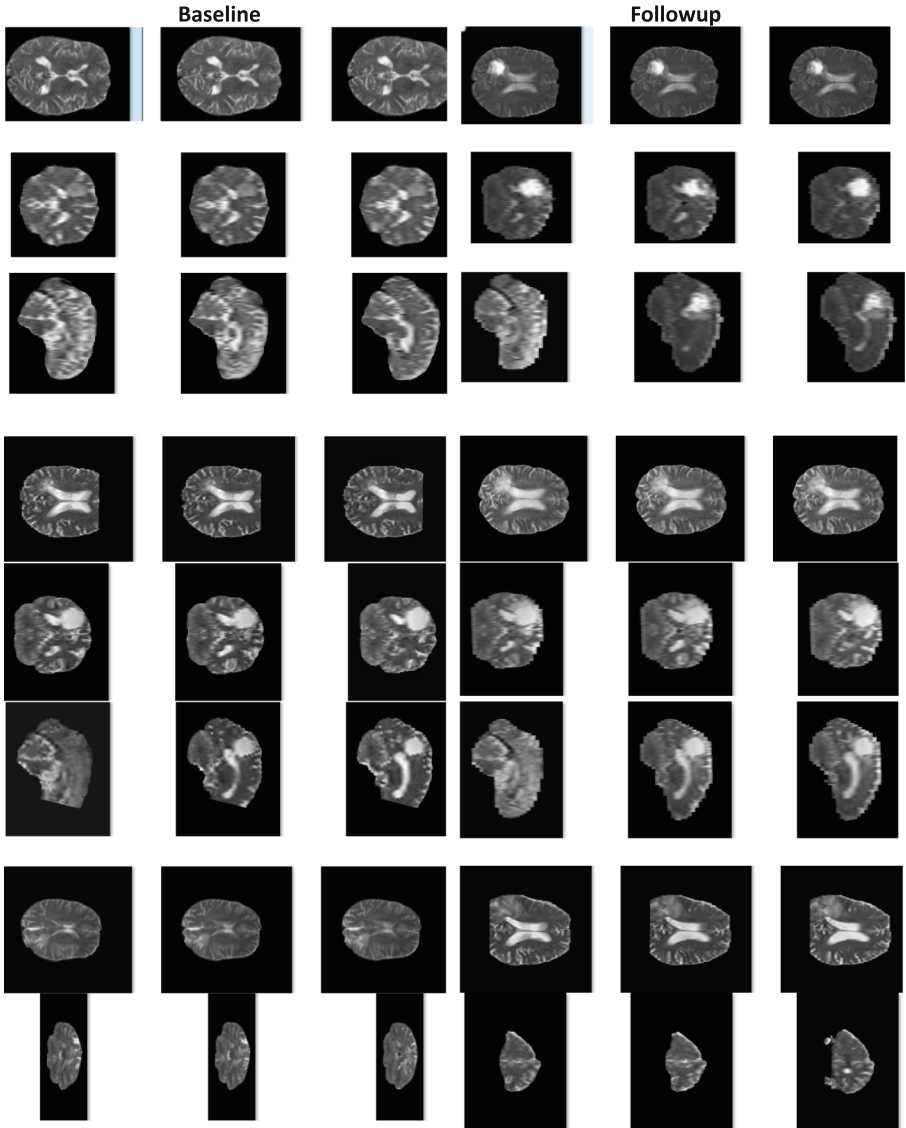


Fig. 5. The illustration of paired sample data for training and testing. The left column is the baseline that applies the Siamese Network 1 in Fig. 1. The right column shows the followup images trained employing the Siamese Network 2 in Fig. 1.

3 Results

From the available datasets from 91 patients, 1432 frames are selected to train the SViT network with a ratio of 80:20 for training and validation. The initial results based on 2D slices appear to be promising with 90% accuracy. The final results are calculated based on all the frames for each subject. For example, if the prediction results of 9 slices are

[2, 3] for 4 classes of [CR, PR, SD, PD] respectively, then the prediction for this patient is PD with the highest probability of [0.22, 0.22, 0.22, 0.33].

If the probabilities of top 2 or more share the same values, a factor of 0.1 is applied to allow the higher weight goes to more serious diseases. This is because of limitation of selections of training and testing 2D slices from 3D volumes. For example, if the probability of 4 classes are [0.3, 0.1, 0.3, 0.3], then the final probability would be [0.3, 0.1, 0.2, 0.4], with the last one (PD) gains 0.1 (+0.1) whereas the neighboring one loses 0.1 (−0.1).

Table 1 provides the final test results based on the test dataset, which ranked **top 2** for this Task 11.

Table 1. The final test results.

	F1	BA	AP
CR	0	0.5	0.0079
PR	0	0.5	0.1847
SD	0.0754	0.4898	0.2663
PD	0.6725	0.4999	0.5123
Mean	0.1870	0.4974	0.2428

4 Conclusion

Due to the limit number of 3D datasets, this SViT network is trained using 2D slices. In future, 3D network will be evaluated to take into account of all voluminous information that a 3D dataset contains.

Acknowledgement. The authors would like to thank the BraTS 2025 organizers for providing not only the valuable dataset but also tireless patient support by answering insightfully all related questions. In addition, the authors are grateful for The Royal Society and The British Council for financial support the early career fellowships (2024–2025) to allow them to participate this competition.

References

1. Suter, Y., et al.: The LUMIERE dataset: longitudinal glioblastoma MRI with expert RANO evaluation. *Sci. Data.* **9**(1), 768 (2022)
2. HD-GLIO-AUTO.: <https://github.com/CCI-Bonn/HD-GLIO-AUTO>
3. DeepBraTumIA.: <https://www.nitrc.org/projects/deepbratumia/>.
4. Kickingeder, P., Isensee, F., et al.: Automated quantitative tumor response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* **20**(5), 728–740 (2019)

5. Guo, J., et al: CMT: convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022)
6. Cho, K., Kim, J., Kim, K., et al.: MuSiC-ViT: a multi-task Siamese convolutional vision transformer for differentiating change from no-change in follow-up chest radiographs. *Med. Image Anal.* **89**, 102894 (2023)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021 (2021)