Multi-Modal Learning for Predicting the Genotype of Glioma

Yiran Wei, *Graduate Student Member, IEEE*, Xi Chen, Lei Zhu, Lipei Zhang, Carola-Bibiane Schönlieb, Stephen Price, and Chao Li

***Abstract*— The isocitrate dehydrogenase (IDH) gene mutation is an essential biomarker for the diagnosis and prognosis of glioma. It is promising to better predict glioma genotype by integrating focal tumor image and geometric features with brain network features derived from MRI. Convolutional neural networks show reasonable performance in predicting IDH mutation, which, however, cannot learn from non-Euclidean data, e.g., geometric and network data. In this study, we propose a multi-modal learning framework using three separate encoders to extract features of focal tumor image, tumor geometrics and global brain networks. To mitigate the limited availability of diffusion MRI, we develop a self-supervised approach to generate brain networks from anatomical multi-sequence MRI. Moreover, to extract tumor-related features from the brain network, we design a hierarchical attention module for the brain network encoder. Further, we design a bi-level multi-modal contrastive loss to align the multi-modal features and tackle the domain gap at the focal tumor and global brain. Finally, we propose a weighted population graph to integrate the multi-modal features for genotype prediction. Experimental results on the testing set show that the proposed model outperforms the baseline deep learning models. The ablation experiments validate the performance of different components of the framework. The visualized interpretation corresponds to clinical knowledge with further validation. In conclusion, the proposed learning framework provides a novel approach for predicting the genotype of glioma.**

***Index Terms*— Multi-modal learning, multi-modal attention, graph neural networks, contrastive learning, brain networks.**

Manuscript received 16 December 2022; accepted 30 January 2023. Date of publication 10 February 2023; date of current version 27 October 2023. This work was supported in part by the National Institute for Health Research (NIHR) (Brain Injury MedTech Co-operative). The work of Stephen Price was supported by the National Institute for Health Research (NIHR), Career Development Fellowship, under Grant CDF-18-11-ST2-003. The work of Chao Li was supported by the Guarantors of Brain Fellowship. *(Corresponding author: Chao Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Local Institutional Review Board.

Yiran Wei, Stephen Price, and Chao Li are with Department of Clinical Neurosciences, Cambridge Biomedical Campus, University of Cambridge, CB2 0QQ Cambridge, U.K. (e-mail: yw500@cam.ac.uk; sjp58@cam.ac.uk; cl674@cam.ac.uk).

Xi Chen is with the Department of Computer Science, University of Bath, BA2 7AY Bath, U.K. (e-mail: xc841@bath.ac.u).

Lei Zhu is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: leizhu@ust.hk).

Lipei Zhang and Carola-Bibiane Schönlieb are with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, CB3 0WA Cambridge, U.K. (e-mail: lz452@cam.ac.uk; cbs31@cam.ac.uk).

Digital Object Identifier 10.1109/TMI.2023.3244038

I. Introduction

GLIOMA is the most common malignant brain tumor in adults with remarkable heterogeneity and diverse survival outcomes [1], [2], [3]. The mutation of the isocitrate dehydrogenase (IDH) gene is one of the most significant molecular markers for the diagnosis and prognosis of glioma [4]. In clinical practice, the most commonly used approaches to determine IDH mutation status, i.e., immunohistochemistry and gene sequencing, rely on tumor samples, which therefore cannot be assessed on those patients who are not suitable for tumor resection or biopsy. Further, as sequencing assays are usually time-consuming and expensive, they are not available in all institutions [4]. Magnetic resonance imaging (MRI) is the mainstay for the management of glioma patients, serving as a routine tool to characterize brain tumors. An increasing number of studies show that MRI can predict the genotype of glioma (e.g. IDH mutations) using machine learning or deep learning models, with a unique non-invasive advantage over the conventional invasive approaches relying on tumor tissue.

Deep learning has achieved better performance than the radiomics approaches based on machine learning models [5]. However, most deep learning models are based on convolutional neural networks (CNN), which cannot leverage the information in non-Euclidean data modalities. For instance, recent studies show that the geometric data describing tumor shape provide robust tumor phenotyping across multiple tissue histology and imaging modalities. In addition, glioma tends to invade the whole brain beyond the focal tumor. Characterizing the global brain using the network approach has shown significance in predicting survival and cognitive decline in brain tumor patients [6], [7]. Hence, integrating these multi-modal data, i.e., tumor image, tumor geometrics, and global brain network, could enhance glioma genotype prediction.

Multi-modal learning shows excellent performance to integrate multiple data modalities while minimizing the domain gap between them. For example, cross-modal attention is shown able to align the fine-grained features between modalities [8]. Additionally, cross-modal contrastive loss shows promising performance in extracting global representations from images and the corresponding texts [9]. Nonetheless, most multi-modal learning methods are designed for the data modalities independent to each other, which may not suit data modalities with geometric relation, i.e., tumor images reflect localized features of focal tumor, while brain networks

contain information from the whole brain. A multi-modal learning scheme which can effectively learn the geometric inter-relation of focal tumor and global brain could better help characterize tumor invasion that widely affects the whole brain.

This study develops a novel learning framework tailored to characterize brain tumor and predict genotype (e.g. binary classification of IDH status: mutations vs wild-types) in glioma. Specifically, apart from the image and geometric data derived from the segmentation masks, we design a self-supervised approach to construct brain networks from anatomical MRIs. Then, we design three separate encoders to extract multi-modal features. In particular, a hierarchical attention is designed to assist the brain network encoder in feature extraction. Afterwards, we design a bi-level multi-modal contrastive loss to tackle the domain gap between the focal tumor and global brain. Finally, we construct a weighted population graph that models the patient cohort based on multi-modal features. A graph neural networks (GNN) is trained to classify patients. Our contributions include:

- Structural brain networks are conventionally constructed from diffusion MRI. To mitigate the limited availability of diffusion MRI, we propose a self-supervised approach of contrastive representative learning to reconstruct the edge attributes of the brain network from anatomical MRI, which could help to mapping the domain knowledge from diffusion MRI to anatomical MRI.
- To allow the brain network encoder extract the most relevant brain network features and reduce the confounding effect from concomitant pathology, we design a hierarchical attention module that sequentially attends to the edges and nodes of the brain network for identifying network features associated with the focal tumor.
- To reflect the gradient tumor invasion and tackle the domain gap across focal tumor and global brain, we present a bi-level contrastive loss, which firstly aligns tumor-level features (i.e., focal tumor image and geometric points cloud) and then further aligns the tumor-level features with the brain-level network features.
- To better integrate multi-modal features and characterise the patient cohort, we construct a population graph for modelling the patient cohort with multi-modal data. The weighted nodes represent the multi-modal features of individual patients, while the weighted edges represent the continuous similarity between patients.

To highlight, we incorporate the domain knowledge of neuroscience and neuro-oncology into the model design. To our best knowledge, this is the first multi-modal learning method based on tumor image, tumor geometrics, and brain networks, to characterize tumor invasion.

II. Related Work

A. Genotype Prediction

The studies of predicting glioma genotypes consist of radiomics-based machine learning methods and deep learning methods. The radiomics-based machine learning approaches first extract hand-crafted features from the tumor core. Feature selection is performed before training models for predicting the genotype [10]. For example, Gihr et al. successfully used intensity-based radiomics features to predict IDH mutation with reasonable accuracy [11]. However, the reproducibility and generalizability of radiomics are often limited by the non-standard feature engineering and selection procedure. The end-to-end deep learning models, i.e., ResNet, DenseNet, provide a more robust prediction for tumor genotype over radiomics approaches [5], [12], [13]. Liang et al. used a 3D-DensNet to predict the IDH mutation, establishing the feasibility of CNN predicting glioma genotype [5]. Other deep learning models incorporate radiomics features into the model. Choi et al. integrated radiomics features into the later layers of CNN to enhance prediction [14], which outperforms the conventional ResNet. Despite achieving reasonable performance, the CNN-based models may not learn the information encompassed in the non-Euclidean data, e.g., geometric points cloud and brain networks, which provide crucial tumor biology and neuroscience information. Hence, we propose specialized encoders to obtain features from multi-modal data.

B. Structural Brain Networks in Glioma

Structural brain network is a graph representation of the complex connectivity among brain regions [15], where the nodes represent the brain regions, defined according to neuroanatomy, and the edges represent the white matter connections among the regions. To generate structural brain networks, most studies use the approaches based on the diffusion MRI, which promises to indicate subtle tumor invasion [7], [16], [17]. However, a robust model training is significantly limited by the data availability of the diffusion MRI. Recent studies indicate that the scalar map of diffusion MRI can be successfully generated from a single anatomical T1 sequence [18], which suggests the high-level correlation between anatomical MRI and diffusion MRI, indicating the potential of constructing brain networks using anatomical MRI. However, a single T1 sequence is insufficient to characterize the heterogeneous structural alternation caused by glioma invasion. Therefore, we proposed reconstructing edge attributes by transferring the knowledge of diffusion MRI to multi-sequence MRI using a contrastive loss. Studies of diffusion-based brain networks generally only include edge attributes. To characterize the brain regions invaded by glioma, we further develop an autoencoder approach to reconstruct node attributes based on regional multi-sequence MRI.

C. Multi-Modal Learning

Multi-modal learning is the deep learning approach that learns from more than one data modality, e.g., images, text, points cloud. Multi-modal learning has shown promising performance in a series of learning schemes. Lee et al. proposed a stacked cross attention to discover the full latent alignments between image regions and words in a sentence. Through inferring image-text similarity, the model produced interpretable prediction results [8]. Zhang et al. employed a

contrastive loss between the lung X-ray and corresponding medical reports to extract relevant representations from both images and text [9]. Nevertheless, existing methods are not designed for data modalities with inclusion relation, e.g., focal tumor and global brain. Therefore, we propose a bi-level contrastive loss to align the features from the focal tumor and global brain levels.

D. Graph Neural Networks

The fast-developing graph neural network (GNN) family promises to extract features and learning from the geometric data, e.g., points cloud, which can be readily reconstructed from MRI [19]. For example, Qi et al. proposed hierarchically generating a graph of points cloud and recursively trained a GNN, which effectively learned local features from the geometric points cloud of the objects.

Further, brain networks are naturally learnable by the GNN due to the graph format. Based on brain networks, GNN has shown high performance in classifying diseases. Ma et al. proposed a combination of recurrent neural network and GNN with an attention-guided random walk module to extract longitudinal structural graph features from the brain network for patient classification [20]. The results showed that the attention mechanism could reveal the most critical brain regions and temporal domain during AD progression. Nonetheless, the attention mechanism designed for other diseases may not suit glioma due to the distinct pathophysiology. We thus develop a hierarchical attention module that could attend to the brain structure to reduce the confounding effect from concomitant pathology and capture tumor-specific features.

Finally, GNN also shows high performance in classifying the nodes in a large graph such as citation networks [21]. The capability of GNN in handling large graphs could be transferred to patient classification tasks. Parisot et al. proposed a population graph to model the dementia cohort by regarding imaging features of individual patients as nodes, while the clinical similarity between patients as edges [22]. A GNN is trained to classify patients, outperforming traditional machine learning models, e.g., random forest. This study develops a population graph to integrate the multi-modal features. Additionally, we permute the edge and node weights to select the best combination in constructing the population graph.

E. Differences From Conference Papers

This study is the extension of our two previous papers in four aspects [23], [24]. Firstly, for brain network reconstruction, we propose a contrastive learning approach to replace the original autoencoder for the brain network edge reconstruction, which additionally incorporates the knowledge from diffusion MRI. Secondly, we combine brain networks with focal tumor data (images and geometrics) to comprehensively characterize glioma. Thirdly, we design an attention module and bi-level multi-modal contrastive loss to extract the most relevant features from the multi-modal data. Finally, we construct a population graph for feature integration and patient classification.

TABLE I
GLOSSARY OF NOTATIONS

Glossary of notations	
Notation	Definition
$x^I; x^P; x^B$	Input image; Input points cloud; Input brain networks
$i; j; k; m; n; M$	Target patient index; Other patient index; Point index in points cloud; Target node index; Other node index; Minibatch size
Multi-modal data generation	
$b^N; b^E; b^{E'}$	Attribute of brain regions/nodes; Attribute of brain edges from anatomical MRI; Attribute of brain edges from FA
$g^E(\cdot); g^{E'}(\cdot);$	Projection head for brain edge attribute extracted from anatomical MRI; Projection head for brain edge attribute extracted from FA;
$z^E; z^{E'}$	Latent feature for anatomical MRI extracted attribute; Latent feature for FA extracted attribute.
$\mathcal{L}_E$	Contrastive loss for extracting brain connection attributes.
Multi-modal contrastive learning	
$f^I(\cdot); f^P(\cdot);$	Image encoder; Geometric encoder
$u^I; u^P; u^B; u^F$	Image features; Geometric features; Brain network features; Focal tumor features
$a^P; a^E; a^N$	Geometric points attention, Brain network edge attention, Brain network node attention
$f^{B'}(\cdot); z^B$	Brain network encoder before pooling; node embedding of brain networks;
$g_a^N(\cdot); g^F(\cdot);$	Projection head of node embedding; Projection head of focal tumor features concatenated from image and geometrics
$g^I(\cdot); g^P(\cdot); g^B(\cdot); g^F(\cdot);$	Projection head for: image features; geometric features; brain network features; focal tumor features
z_m^N	Node embedding of mth node in brain networks
$z^I; z^P; z^B; z^F;$	Latent feature of: image; geometrics; brain networks; focal tumor
$l^{I2P}; l^{P2I}; l^{B2F}; \mathcal{L}_{multi}$	Contrastive loss from image to geometrics; Contrastive loss from geometrics to image; Contrastive loss from brain networks to focal tumor; Multi-modal contrastive loss
Population graph	
$w^{edge}; w^{node}$	Edge weight of population graph; Node weight of population graph

III. METHODOLOGY

A. Study Overview

The proposed prediction model includes three stages: (Fig. 1): (1) generating multi-modal data of tumor image, tumor geometrics and brain networks from the multi-sequence MRI; (2) multi-modal contrastive learning extracting features from focal tumor image, geometrics and global brain networks; (3) feature integration to construct a population graph for patient classification and genotype prediction.

In this study, we introduced two types of graph data: brain network and population graph. To distinguish the features extracted from the two types of graphs, we define the features from brain networks as attributes and features from the population graph as weight. The superscript E/N only denotes the brain network edge/node in this study. In addition, node embedding represents the aggregated node features during graph convolution of brain networks. We present the glossary of notations in Table I.

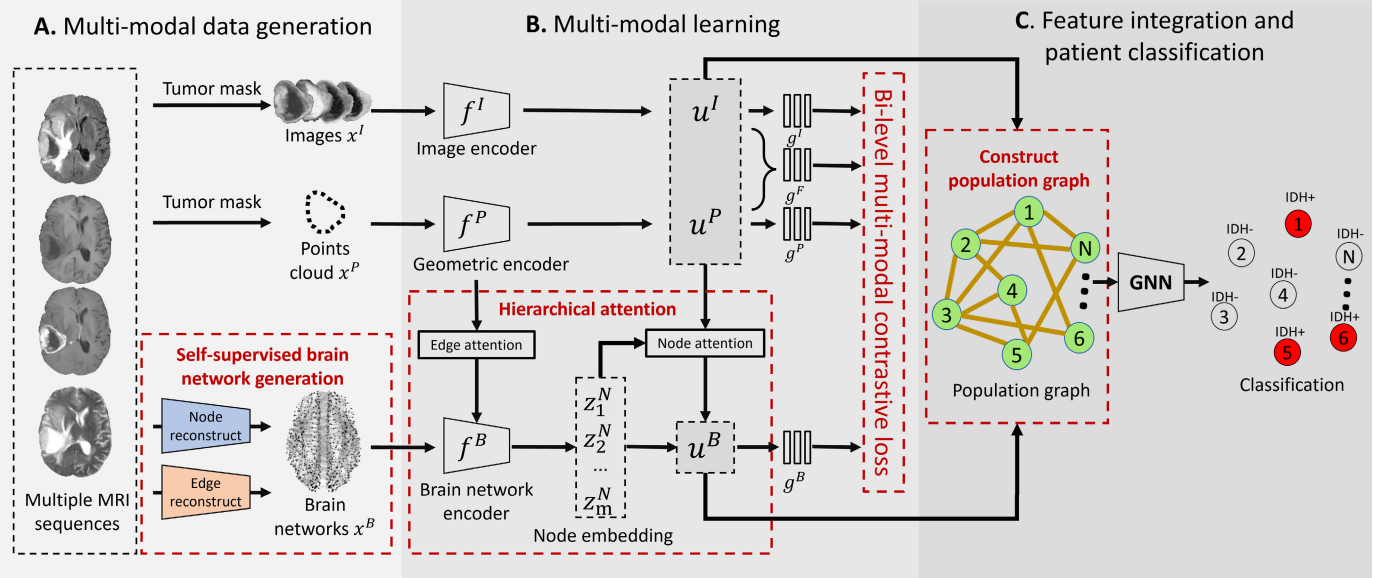

Fig. 1. Study overview: A. Multi-modal data generation. Image x^I and geometric x^P data are generated from tumor masks, while brain network data x^B are generated from the pretrained self-supervised models. B. Features of tumor image (u^I), tumor geometrics (u^P), focal tumor($u^F = \langle u^P, u^I \rangle$) and global brain network (u^B) are projected by respective projection heads (g^I, g^P, g^F and g^B) for bi-level contrastive learning. A hierarchical attention module attend to the edges and nodes in the brain network. C. A population graph is used to integrate multi-modal features and classify patients using a GNN.

B. Multi-Modal Data Generation

Our method starts by generating three data modalities from the input multiple MRI sequences (see Fig. 1A), and the three data modalities are: (1) the image data of focal tumor (denoted as x^I) is obtained by assigning Boolean values on the tumor masks and the MRI; (2) the tumor geometric data (denoted as x^P), in the form of points cloud, is generated by sampling the surface meshes of tumor masks using a standard farthest point sampling strategy; and (3) the brain networks (denoted as x^B) are generated by two self-supervised neural networks (NNs) detailed below.

1) Brain Networks Construction Via Self-Supervised NNs: The brain networks, consisting of reconstructed nodes and edges, are generated based on a prior neuroanatomy atlas.

In this study we denote the edges and nodes of brain networks as attributes of edges b^E and attributes of nodes b^N. The brain node attribute is defined as the features extracted by the pre-trained node autoencoder from the anatomical MRI regions defined by the anatomical atlas. The node attribute represents the independent regional features of the separated brain areas. The brain edge attribute is defined as the feature extracted by the pre-trained edge encoders from the anatomical MRI representing the white matter tract connectivity across different brain regions defined by the anatomical prior. As shown in the upper half of Fig 2, the anatomical MRI is the only input to the autoencoder for reconstructing the node attributes at both the training and testing phases. The trained encoder extracts node attributes from the prior brain atlas (as the nodes of the brain network). Specifically, voxels enclosed by the 90 cortical/ subcortical brain regions on the atlases [25] are extracted and fed into an NN-based autoencoder (AE) to produce the brain node attributes a^N of the brain networks. The AE consists of a NN encoder that extracts high-level

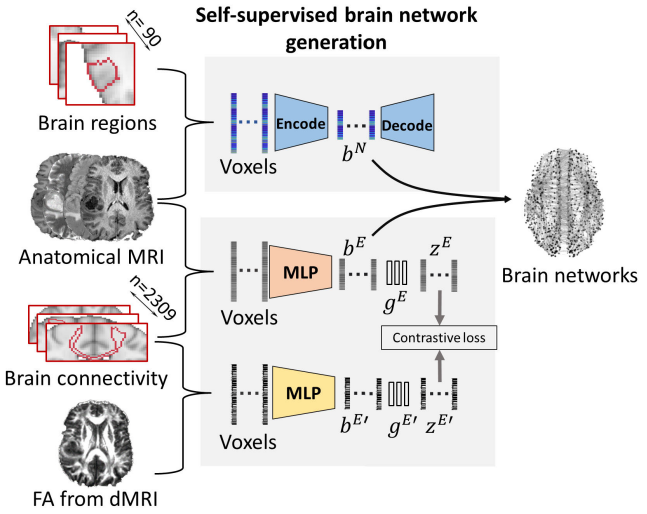

Fig. 2. Brain network generation. Two self-supervised models are trained to extract node/edge attributes (b^N, b^E) from node/edge atlas bounded MRI voxels: Node attributes are extracted by the autoencoder, while edge attributes are reconstructed through contrastive learning between anatomical MRI and FA map of dMRI using projection head (g^E, $g^{E'}$), projected latent features (z^E, $z^{E'}$) and a contrastive loss $\mathcal{L}_E$.

representation vectors from the voxels in the brain node and a NN decoder that attempts to restore the voxels from the representation vectors. By adopting this self-supervised model, representations of the voxels in the brain regions could be extracted as brain node attributes. In contrast, the input to the contrastive learning encoders (MLP) for the edge attributes includes both FA maps and anatomical MRI in the training stage, whereas the FA maps and the corresponding encoder are no longer needed in the testing stage. The trained encoder of anatomical MRI extracts edge attributes from the tract atlas (as the edges of the brain networks). Particularly, we use the tract atlas as the regions of interest for reconstructing brain edge

attributes b^E. Reference [16], indicating the 2,309 pathways of white matter tracts connecting the 90 brain regions. Due to the clinical significance of the fractional anisotropy (FA) map derived from the diffusion MRI in characterizing brain connectivity/edge, we utilize the FA map to guide the attributes extraction of anatomical MRI. Firstly, voxels of anatomical MRI and the corresponding FA map enclosed by the tract atlas are input into two multilayer perceptron (MLP), which respectively extract the attribute vectors b^E and $b^{E'}$ from voxels. Next, two projection heads $g^E(\cdot)$ and $g^{E'}(\cdot)$ project the attributes to a common latent space, where domain alignment is performed between the latent attributes of anatomical MRI ($z^E = g^E(b_n^E)$) and FA ($z^{E'} = g^{E'}(b_n^{E'})$) using a contrastive loss. The brain edge attributes extracted from the anatomical MRI would contain corresponding information in the FA map. The contrastive loss for the brain edge $\mathcal{L}_E$ is defined as:

$$\mathcal{L}_E = \frac{1}{M}\sum_{i=1}^{M}(-\log\frac{\exp(S(z_i^E, z_i^{E'})/\tau)}{\sum_{j\neq i}^{M}\exp(S(z_i^E, z_j^{E'}))/\tau)}), \quad (1)$$

where i is the index of target patients, while j is the index of other patients in the minibatch M; $S(\cdot)$ is the similarity score; τ is the temperature parameter setting to 0.1; M is the size of the minibatch. The negative pair is the edge attribute pair (FA, anatomical MRI) of different patients, whereas the positive pair is the edge attribute pair of the same patient. The loss maximizes the distance between the positive pair and minimizes the distance between the negative pair to extract the most FA-relevant features from anatomical MRI. The node and edge attributes are independent features linked together according to the prior anatomical connection for constructing the brain network: $x^B = \{b^E, b^N\}$.

C. Multi-Modal Learning for Image, Geometrics and Brain Networks

The proposed multi-modal learning framework extracts features from the three modalities of data, i.e., focal tumor image, focal tumor geometric, and global brain networks. Moreover, hierarchical attention is developed for the brain network encoder to extract tumor-related brain network features. Finally, the extracted features are projected into a shared latent space for bi-level multi-modal contrastive learning, which could minimize the domain gap from the tumor level (image and geometrics) across the global brain level (focal tumor and brain networks). As shown in Fig. 1B, the projection is conducted via three NN-based encoders as follows.

1) Image Encoder: The image encoder is a 3DCNN defined by $u_i^I = f^I(x_i^I)$, where x_i^I and u_i^I are the image data and output features for the ith patient, and $f^I(\cdot)$ is the 3DCNN model (see Section IV-C for implementation details).

2) Geometric Encoder: $f^P(\cdot)$ is defined as the geometric encoder (Fig. 3A) outputs the geometric features u_i^P and geometric attention $a_{i,k}^P$ for kth point in the points cloud, defined as $u_i^P, a_{i,k}^P = f^P(x_i^P)$ for the ith patient.

3) Brain Network Encoder With Hierarchical Attention: Brain network features are extracted by training a NN with graph convolution layers, where the NN weights are corrected following a novel hierarchical attention mechanism.

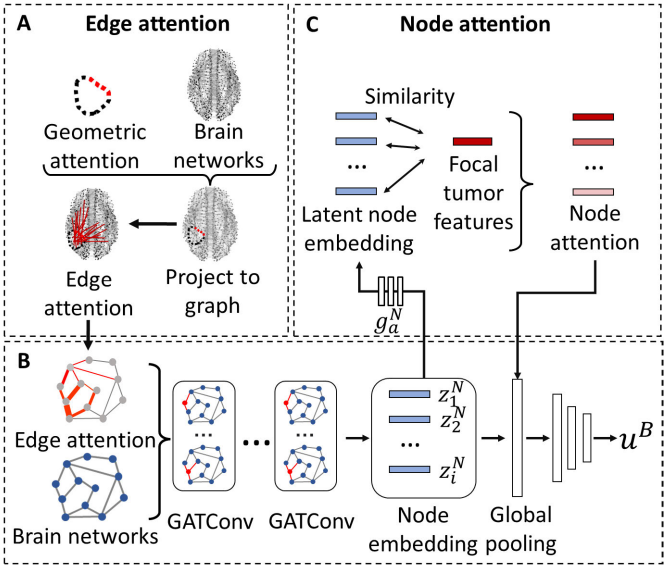

Fig. 3. Hierarchical graph attention: A. Geometric boundary attention produced by the geometric encoder is projected to brain networks to obtain edge-level attention. B-C. Edge-attended brain networks are convoluted to produce node embeddings, projected to latent space for generating node-level attention by computing similarity with tumor features. The node-level attention is then utilized in the global pooling level for generating tumor-related brain-level network features u^B.

The attention mechanism is structured by edge-level attention and node-level attention. The former is obtained by projecting the geometric attention of tumor boundary onto the edges (Fig. 3B). Specifically, the points clouds are projected to the edge atlas. The crossing edges are then assigned with the boundary attention of the points cloud. The edge attention is defined as:

$$a_{i,(m,n)}^E = \frac{1}{K}\sum_{k}^{K}(a_{i,k}^P), \quad (2)$$

where $a_{i,(m,n)}^E$ is the edge attention of (m,n)th edge connecting mth node and nth node of the ith patient. K is the number of points in points cloud crossed by (m,n)th edge and $a_{i,k}^P$ is the attention of kth point crossed by the (m,n)th edge.

The outputs of the edge-level attention are further encoded by the GATConv layers that convolute the nodes and edges of the brain networks to obtain the node m embedding in patient i defined by $z_{i,m}^N = f^{B'}(x_i^B)$, where $f^{B'}$ is the components of brain network encoder before the global pooling layers (Fig. 3B). Afterwards, the node embeddings are projected to the latent space by a projection head g_a^N. To extract the tumor-related node embeddings, we applied another projection head g^F to project the concatenated focal tumor features of ith patient: $u_i^F = \langle u_i^I, u_i^P \rangle$, composed by both images and points cloud, into the latent space shared with node embedding. We measure the similarity between the mth node embedding with the tumor features of ith patient by:

$$a_{i,m}^N = S(g_a^N(z_{i,m}^N), g^F(\langle u_i^I, u_i^P \rangle)), \quad (3)$$

where $a_{i,m}^N$ is the attention of the mth node of ith patient. $z_{i,m}^N$ is the node embedding of mth node of ith patient. $g^F(\cdot)$ and $g_a^N(\cdot)$ are projection heads projecting tumor features

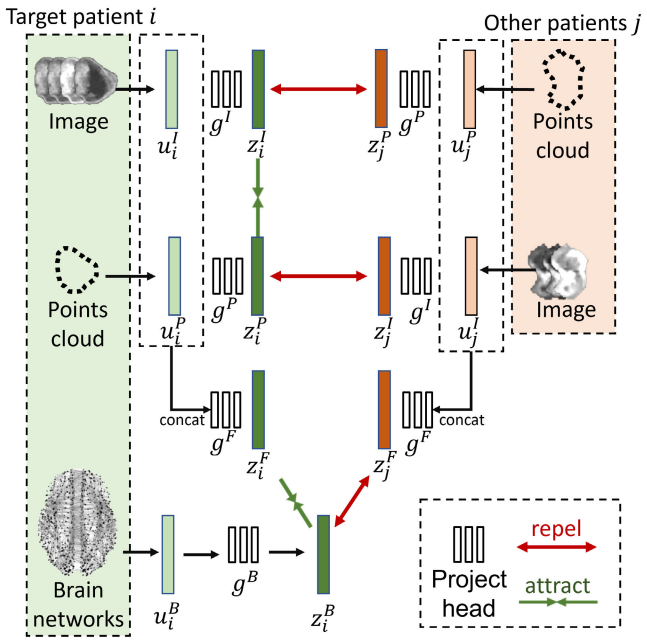

Fig. 4. Bi-level multi-modal contrastive learning: Latent features of different modalities (u_i^I, u_i^P), (u_i^F, u_i^B) from the same patient (green) attract each other, while latent features of different modalities (u_i^I, u_j^P), (u_j^F, u_i^B) from different patients (red) repel each other. The bi-level loss consists of a tumor-level and brain-level components trained together.

(e.g. concatenated image and geometric features $\langle u_i^I, u_i^P \rangle$) and node embeddings to the same latent space; $S(\cdot)$ is the similarity function.

By performing attention in training the brain network encoder, we extract the most tumor-related features from the brain network and reduce the noise caused by confounding effects, e.g., ageing or other concomitant pathology (Fig. 3B). The feature extraction of the brain networks is defined as $u_i^B = f^B(x_i^B)$, where x_i^B and u_i^B are the brain network data and brain network features for the ith patient, and f^B represents the GNN-based brain network encoder.

4) Bi-Level Multi-Modal Contrastive Loss: We develop a bi-level multi-modal contrastive loss to further characterize tumor gradient invasion and minimize the domain gap between the focal tumor and global brain. After extracting the multi-modal features from different encoders, two projection heads are adopted to respectively project the tumor-level features of images and points cloud to the same latent space: $z_i^I = g^I(u_i^I)$, $z_i^P = g^P(u_i^P)$ where z_i^I and z_i^P are the projected latent features of images and points cloud, g^I and g^P are the pre-defined projection heads.

Meanwhile, another two projection heads are employed to respectively project the extracted focal tumor features and brain network features into another latent space: $z_i^B = g^B(u_i^B)$, $z_i^F = g^F(u_i^I, u_i^P)$, where z_i^B and z_i^F are the projected latent feature of brain networks and focal tumor); g^B and g^F are the projection head for the brain network and focal tumor.

Subsequently, a bi-level multi-modal contrastive loss is developed to firstly reduce the domain gap of tumor-level features by minimizing the cosine distance (attract) between the multi-modal latent features (z_i^I, z_i^P) from the same patient i and maximizing the cosine distance (repel) of multi-modal latent feature pairs (z_i^I, z_j^P), (z_i^P, z_j^I) from different patients i and j using the contrastive loss. Secondly, the brain-level domain gap is optimized using a similar approach for the features of brain networks (z_i^B) and focal tumor (z_i^F). We design three contrastive losses for tumor image to tumor geometrics (Equation. 4), tumor geometrics to tumor image (Equation. 5) and global brain network to focal tumor (Equation. 6). Finally, we integrate those three sub-losses with a weighting coefficient λ.

$$l_i^{I2P} = -\log \frac{\exp(S(z_i^I, z_i^P)/\tau)}{\sum_{j\neq i}^{M} \exp(S(z_i^I, z_j^P)/\tau)}, \tag{4}$$

$$l_i^{P2I} = -\log \frac{\exp(S(z_i^P, z_i^I)/\tau)}{\sum_{j\neq i}^{M} \exp(S(z_i^P, z_j^I)/\tau)}, \tag{5}$$

$$l_i^{B2F} = -\log \frac{\exp(S(z_i^B, z_i^F)/\tau)}{\sum_{j\neq i}^{M} \exp(S(z_i^B, z_j^F)/\tau)}, \tag{6}$$

where i is the index of the target patient, and j is the index of other patients in the mini-batch; $S(\cdot)$ is the similar score function; τ is the temperature parameter (setting to 0.1) controlling strength of penalties on negative examples. The positive pairs are defined as the modality pairs from the same patient, while negative pairs are defined as the modality pairs from different patients. M is the size of the mini-batch. l^{I2P} is the image to points cloud contrastive loss. l^{P2I} is the points cloud to image contrastive loss. l^{B2F} is the brain network to focal tumor contrastive loss. It's necessary to define different losses for different pairs of modalities because contrastive loss is asymmetric for each input modality [9].

$$\mathcal{L}_{multi} = \frac{1}{M}\sum_{i=1}^{M}(\lambda(\frac{l_i^{P2I} + l_i^{I2P}}{2}) + (1-\lambda)l_i^{B2F}), \tag{7}$$

where $\lambda \in [0, 1]$ is a scalar weight coefficient. M is the minibatch size. The final multi-modal contrastive loss $\mathcal{L}_{multi}$ is computed as a weighted combination of three contrastive losses that maximizes the distance between multi-modal features of the different patients and minimizes the multi-modal features of the same patient.

5) The Algorithm: The proposed multi-modal contrastive learning algorithm is shown in Algorithm 1.

D. Population Graph for Classifying Glioma Patients

With the focal tumor and brain network features generated from the multi-modal learning, we construct a population graph to characterize the patient cohort (Fig. 1C): each node represents the multi-modal features extracted from the patients, while each edge represents the similarity between the multi-modal features among the patients (Fig. 5). In the population graph, the node weight of patient i is defined as $w_i^{node} = u_i$, and the edge weight between patient i and j is defined as:

$$w_{i,j}^{edge} = \begin{cases} r(u_i, u_j), & \text{if } r(u_i, u_j) \geq \theta \\ 0, & \text{if } r(u_i, u_j) < \theta \end{cases} \tag{8}$$

where $u \in \{u^F, u^B, \langle u^F, u^B \rangle\}$: u is the feature extracted from the multi-modal contrastive learning, and $r(\cdot)$ is the correlation

Algorithm 1 Multi-Modal Contrastive Learning

Input: image: x_i^I, points cloud: x_i^P, brain network: x_i^B
for $i = 1, \cdots 270$ **do**
  Compute features and attention from image and geometric points cloud: $u_i^I = f^I(x_i^I)$;
  $u_i^P, a_{i,k}^P = f^P(x_i^P)$;
  Input edge attention: $x_i^{B'} = x_i^B \cdot a_{i,(m,n)}^E$ via (2).
  Compute node embedding using brain network encoder: $z_{i,m}^N = f^{B'}(x_i^B)$.
  for $m = 1, \cdots 90$ **do**
    Compute node attention $a_{i,m}^N$ via (3).
  end for
  Extract features from brain networks: $u_i^B = f^B(z_i^N \cdot a_{i,m}^N)$.
  Project features to latent space:
  - Image: $z_i^I = g^I(u_i^I)$
  - Points cloud: $z_i^P = g^P(u_i^P)$
  - Focal tumor: $z_i^F = g^F(\langle u_i^I, u_i^P \rangle)$
  - Brain networks $z_i^B = g^B(u_i^B)$

  Compute multi-modal contrastive loss by (7).
end for

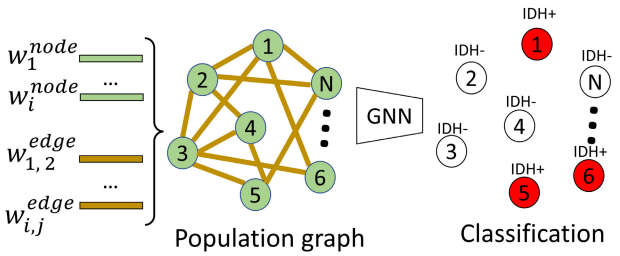

Fig. 5. Population graph for patient classification: Each node weight (w^{node}) represents features of one patient, while each edge weight (w^{edge}) represents the similarity among the features of patients. A GNN node classifier is trained for classifying patients.

operator. θ is the threshold of the correlation. We design five different combinations of the node weight w^{node} and the edge weight w^{edge} listed in Table II.

Specifically, we consider different combinations of focal tumor features (tumor image and geometrics) and global brain network features as edge and node weight, and we construct the population graph based on the hypothesis that the two categories of features may reflect different patterns of tumor invasion, i.e., localized v.s. widespread invasion. As such, we could integrate two types of features and characterize both the homogeneity and heterogeneity of the cohort.

IV. Experiments

A. Datasets

We collect the anatomical MRI data of 407 glioma patients available from The Cancer Imaging Archive (TCIA) [26], [27], [28] and an in-house cohort. The cohort includes 105/407 (25.7%) IDH mutants and 302/407 (74.3%) IDH wild-types. The MRI modalities include pre-contrast and post-contrast T1, T2, and T2-FLAIR.

A total of 20 patients of the in-house cohort is used for training the self-supervised models of brain network construction.

TABLE II
Population Graph With Different Node and Edge Weights

Node weight	Edge weight
u^F	$r(u^B)$
u^B	$r(u^F)$
$\langle u^F, u^B \rangle$	$r(u^B)$
$\langle u^F, u^B \rangle$	$r(u^F)$
$\langle u^F, u^B \rangle$	$r(\langle u^F, u^B \rangle)$

The remaining 387 patients are split into training and testing set with a 7:3 ratio. The training set of 270 patients is divided by half to train the bi-level multi-modal contrastive learning for feature extraction and population graph-based classifier for patient classification. The testing set includes 117 patients from the publicly available TCIA website. For baselines, we used the training data of 270 patients to train the models. We then evaluated the trained models with independent testing data of 117 patients.

B. Image Pre-Processing

A standard pre-processing pipeline on MRI data is performed as described [29]. Firstly, the pre-contrast T1, T2, and FLAIR images are co-registered to the post-contrast T1 images using the FMRIB's Linear Image Registration Tool of the FMRIB Software Library (FSL) [30]. Next, skull stripping is performed using the Brain Extraction Tool in FSL [31]. Finally, histogram matching [32] and voxel smoothing with SUSAN noise reduction [30] are conducted as normalization.

For the in-house cohort with diffusion MRI modalities available, the FA maps are derived from the diffusion MRI using the FMRIB's Diffusion Toolbox. The FA maps are used to train the self-supervised models to extract tract-related features from the anatomical MRI to generate brain networks.

Finally, all the MRI data are non-linearly transformed to the standard space by co-registering them to the MNI-152-T1-2MM-brain template available in the FSL using the Advanced Normalization Tools (ANTs) [33]. ANTs is a commonly used tool for the co-registration of the lesion-bearing brain registration, which is shown to mitigate the distortion caused by brain tumor mass effect [34].

The tumor mask contains the tumor core defined by the brain tumor segmentation benchmark. Reference [35]. The tumor mask is generated from the following steps: we first generate an initial tumor segmentation (auto-mask) using a pre-trained nnUNet [36], which is then inspected by two clinical experts. The segmentation errors of the initial results are manually corrected. The final resulting mask is then taken as the ground-truth segmentation mask used in our experiments. We follow the standard annotation procedures of the multimodal brain tumor segmentation (BraTS) benchmark reported by [35]

C. Implementation Details

1) Brain Network Generation: The 20 patients yield 46,180 edges and 1,800 nodes for training the self-supervised model.

For the node autoencoder, all input node voxels are sampled to the dimension of 4,000, the encoder consists of six layers

(dimension 2048, 1024, 512, 128, 32, 16), and the output of the bottleneck is the attribute feature b^N with a dimension of 16. For the edge encoder, all input edge voxels are sampled (Anatomical: 4000, FA: 1000). Two MLP (MRI: 2048, 1024, 512, 128, 32, 16; FA: 1024, 512, 128, 32, 16) respectively encode the voxel vectors of T1 and FA to attribute vectors b^E and $b^{E'}$ with a dimension of 16. The final brain networks contain 90 nodes and 2,309 edges with a dimension of 16.

2) Image Encoder: The image encoder is a 3DCNN architecture consisting of five 3D convolutional layers (dimension 64, 128, 128, 256, 256), with four input channels corresponding to the four MRI sequences. Batch normalization and max pooling are performed for all convolutional layers. Three feed-forward layers (dimension: 512, 256, 32) are followed to output features with a dimension of 32.

3) Geometric Encoder: A specialized GNN is adopted to extract features from the points cloud. The points are first converted into a graph for each convolution by generating links between points and their nearest neighbors within a predefined radius distance. Secondly, convolution operators NNConv [37] aggregate the points features (euclidean coordinates of points) and the link features (distance between points) to the center node. Finally, the farthest points sampling is adopted to sample the points with the furthest distance from other points. Our geometric encoder consists of convolutional layers (dimension: 32, 64, 64, 128, 128). After the last layer, a global attention pooling is employed to produce attention scores for the points. Finally, a feed-forward network (dimension: 256, 128, 32) outputs the geometric features with a dimension of 32.

4) Brain Network Encoder: The projection heads g_a^N for node attention is NNs (dimension: 16, 32, 64, 128) that projects node embeddings to the latent space shared with focal tumor features z^F. The brain network encoder is a graph attention network with GATConv layers (dimension: 64, 128, 128, 256, 256, 256) that can handle the high-dimensional node and edge attributes [38]. The feed-forward network outputs the brain network features (dimension: 512, 256, 32). Cosine similarity is used as the similarity score $S(\cdot)$ for generating node attention.

5) Bi-Level Multi-Modal Contrastive Loss: The projection heads g^I, g^P and g^B, are three separate NNs (dimension 32, 64, 128, 128) that project the features to the latent space with dimension of 128. The projection head for g^F is another NN (dimension 64, 64, 128, 128) that projects the $\langle u^I, u^P \rangle$ to the latent space with a dimension of 128. Cosine similarity is selected as the similarity score $S(\cdot)$. τ is set to 0.1 λ is set to 0.8.

6) Population Graph and GNN Classifier: θ of the population graph is set to 0.5. The GATConv is employed as the graph kernel of the GNN to perform node classification in the population graph. The GNN for the population graph consists of layers of GATConv (dimension: 64, 128, 128, 128) followed by pooling layers and a classification layer.

7) Training Parameters: For self-supervised learning for generating brain networks, the autoencoder adopts mean squared error loss (MSELoss), the Adam optimizer with a weight decay of 0.0005 and a batch size of 50. We implement the following hyperparameters: 1000 training epochs. We set the initial learning rate as 0.001, and the learning rate is reduced to 90% after every 50 epochs. For edge reconstruction, we adopt the SGD optimizer [39] with a weight decay of 0.0005 and a batch size of 50 with 1000 training epochs. We set the initial learning rate as 0.001, and the learning rate is reduced to 90% after every 50 epochs.

For multi-modal learning, we adopt the SGD optimizer to optimize the network with a weight decay of 0.0005 and a batch size of 20. We implement the following hyperparameters: 1000 training epochs; a mini-batch size of 20. We set the initial learning rate as 0.001, and the learning rate is reduced to 90% after every 50 epochs. Data augmentation is performed by rotating both images and points cloud data with the same angles.

For population graph-based-GNN, we adopt the Adam optimizer [39] with a weight a batch size of 20. We apply binary cross-entropy loss for patient classification. We implement the following hyperparameters: 200 training epoch; a mini-batch size of 20. We set the initial learning rate as 0.001, and the learning rate is reduced to 90% after every 50 epochs. We used a semi-supervised training approach proposed in [22] and [40] to perform node classification. During training, both the training and testing data are included in the large graph, but only the labels of training data are available. Once trained, the graph is applied to classify testing data.

To avoid over-fitting, we applied early stopping, drop-out layers, and regularization during training, which stop training once the model performance stops improving on a hold out validation dataset for 5 epochs.

D. Model Evaluation

1) Evaluating Performance of the Overall Framework: To evaluate the overall performance of our approach for predicting IDH status, we implement four published methods as the benchmark including a ResNet-based sequence network proposed by Chang et al. [12], a DenseNet based network proposed by Liang et al. [5], a hybrid model using ResNet with integrated radiomics features proposed by Choi et al. [14] and radiomics feature based approach proposed Peng et al. [41]. In addition, ResNet34 and DenseNet50 backbones are also applied to perform classification on both tumor-only images and whole-brain images respectively. Finally, we implemented another multi-modal learning framework DCL-NET proposed by Lin et al. [42] as the baseline for the multi-modal learning. We modified the DCL-NET to suit our task by substituting the three inputs including two views and one points cloud with images, points cloud and Brain networks. In addition, the encoders are substituted accordingly while IDH-mutant and IDH wild-types are regarded as two labels for classification. All baselines are trained with the binary cross-entropy loss. To evaluate the performance of population graph comparing to other traditional machine learning models. We also implement the MLP and support vector machines (SVM) as the benchmarks to compare the proposed population graph-enhanced GNN. The performance of models are evaluated using the area under the curve (AUC) of the receiver operating characteristic, accuracy (ACC), specificity (SPE) and sensitivity (SEN).

2) Evaluate Population Graphs: We conduct experiments in constructing a population graph to choose the best combination of edge and node weights (Table II).

3) Ablation Experiments: We perform ablation experiments to test the importance of different components in the proposed framework. Specifically, we first test the performance of every single encoder of tumor image, tumor points cloud and brain networks, using an MLP and binary cross-entropy loss. Secondly, we test the performance of the pairwise combinations of training two encoders for the classification. Thirdly, we add the contrastive loss L_{contra} to the training of the above two-encoder combinations for multi-task experiments. All of above experiments test the importance of each modalities in the multi-modal input. Further, we implement a multi-modal framework without the contrastive loss and a multi-modal framework without the hierarchical attention to evaluate the importance of those two components in the multi-modal learning. Finally, we substitute the population-graph-based GNN with a MLP end to test the usefulness of the population graph.

E. Visualize Interpretation

To interpret the results of the proposed multi-modal learning framework, we identify the critical regions contributing to the prediction from tumor images, tumor points cloud and brain networks using different interpretation approaches.

We employed the Grad-CAM [43] to visualize the critical regions on the tunor images and visualize the geometric attention of the points cloud. To visualize the concordance between points cloud and tumor image, we project the surface points of the Grad-CAM map overlaid on the tumor image to the corresponding points cloud.

To interpret the learning process of the brain network encoder, we employ the GNNExplainer [44] to output a probability score that infers the importance of the edges in the brain network. We retain those edges with probability scores greater than 50%.

V. Results

A. Performance of Population Graph

The performance of different approaches of constructing population graph is in Table III. The results show that the population graph achieves the best performance (AUC 0.962) with the concatenated tumor features and brain network features defined as the node and the cosine similarity between brain network features defined as edge. The population graph that uses similarity between tumor features to define edge ($u^B, r(u^F)$: AUC 0.914, $\langle u^F, u^B\rangle, r(u^F)$: AUC 0.940) generally performs worse than those using the similarity between brain network features to define edge ($u^F, r(u^B)$: AUC 0.939, $\langle u^F, u^B\rangle, r(u^B)$: AUC 0.962). Strikingly, the population graph with concatenated tumor features and brain network features defined as both node and edge performs the worst (AUC 0.888).

B. Performance of the Proposed Framework

Our results (Table IV) show that the best setting of multi-modal framework (AUC 0.962) outperforms published baselines (Radiomics: AUC 0.743, ResNet-based: AUC 0.858, DenseNet-based: AUC 0.888, ResNet + Radiomics: AUC 0.822, DCL-NET: AUC 0.897) the CNN backbones (DenseNet + Tumor: AUC 0.938, ResNet + Tumor: 0.907, DenseNet + Brain: AUC 0.719, ResNet + brain: 0.711). It is worth mentioning that the CNN-backbone models with whole brain image input had much worse performance than the models with tumor image input, which supports the usefulness of our multi-modal learning approach and the proposed attention module in extracting robust features from the whole brain image. The better performance of our approach comparing to the published multi-modal framework (DCL-NET) also suggests the usefulness of the hierarchical attention and multi-modal loss designed for the brain networks. Notably, the performances of 3D-CNN models are higher than the combination of multi-modal contrastive learning with traditional machine learning models (MLP: AUC 0.936, SVM: AUC 0.932), implying the importance of the population graph for feature integration.

TABLE III
Experiment for Selecting the Best Combination of the Population Graph

Node weights	Edge weights	AUC	ACC	SEN	SPE
u^F	$r(u^B)$	0.939	0.879	0.894	0.812
u^B	$r(u^F)$	0.914	0.836	0.833	0.843
$\langle u^F, u^B\rangle$	$r(u^B)$	**0.962**	**0.905**	**0.917**	**0.875**
$\langle u^F, u^B\rangle$	$r(u^F)$	0.940	0.862	0.857	**0.875**
$\langle u^F, u^B\rangle$	$r(\langle u^F, u^B\rangle)$	0.888	0.862	0.905	0.750

TABLE IV
Comparing With Benmarks

Models	AUC	ACC	SEN	SPE
Published baselines				
Radiomics [41]	0.743	0.724	0.738	0.688
ResNet-based [12]	0.858	0.795	0.798	0.781
DenseNet-based [5]	0.888	0.819	0.833	0.781
ResNet + Radiomics [14]	0.822	0.793	0.810	0.750
DCL-NET [42]	0.897	0.819	0.857	0.719
CNN backbones				
ResNet+Tumor	0.907	0.810	0.809	0.813
ResNet+Brain	0.711	0.647	0.607	0.750
DenseNet+Tumor	0.938	0.853	0.857	0.844
DenseNet+Brain	0.719	0.672	0.655	0.719
Baselines for GNN				
Multi-modal + MLP	0.936	0.862	0.893	0.781
Multi-modal + SVM	0.932	0.871	0.893	0.813
Proposed framework	**0.962**	**0.905**	**0.917**	**0.875**

C. Ablation Experiments

The full results of the ablation experiments are shown in Table V. For multi-modal input experiments: the experiments of the individual encoder show that the brain encoder (AUC 0.877) outperforms the tumor geometric (AUC 0.858) and tumor image (AUC 0.869) encoders. The multi-task two-encoder experiments indicate that the best combination of data modalities is tumor image and tumor points cloud (AUC 0.874). The two-encoder experiments with the additional contrastive loss show that combining tumor image and tumor geometric encoders (AUC 0.929) consistently performs the best. The contrastive loss significantly improves the two-encoder setup. For the multi-modal learning stage,

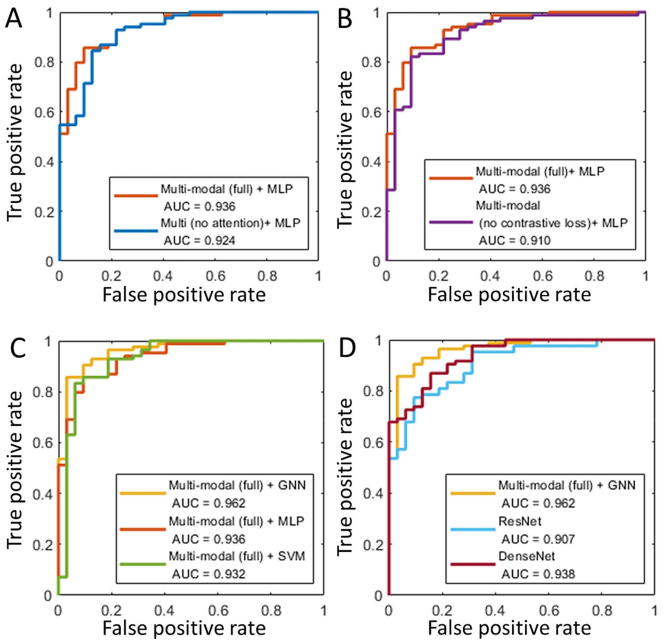

Fig. 6. AUC for the ablation experiments: A. Models with and without hierarchical attention. B. Models with and without bi-level contrastive loss. C. Models of the full framework and the benchmarks of population graph-enhanced GNN. D. Models of the full framework and CNN benchmarks.

removing the graph attention modules significantly decreases the performance of the contrastive training framework (with attention: AUC 0.936, without attention: AUC 0.924). The model without multi-modal contrastive loss also generates a lower performance (with L_{multi}: AUC 0.936, without L_{multi}: AUC 0.910).

Further, we evaluate the importance of the population graph: the replacement of population-graph-based GNN with MLP will result in a lower AUC (MLP: AUC 0.935, Population graph-based GNN AUC 0.962). Finally, we test the model performance with and without manually corrected tumor mask, results showed that our framework slightly relies on the accurate mask, however the framework with the automatically generated mask still outperforms published models in Table IV.

D. Interpretative Visualization

The interpretative visualizations in Fig.7 show that the image encoder and the geometric encoder indicate common tumor regions (Fig.7D - F) important for model prediction, which suggests that these regions are specific to IDH mutation. Fig.7D shows that the Grad-CAM focuses on the tumor contrast-enhancing edges with high intensity in both T2 and T2-FLAIR images (Fig.7B, C). Through the visualization in Fig.8, we find that the brain networks of IDH wild-type demonstrate a higher density of important disrupted edges compared to IDH mutant. This finding aligns with our prior knowledge that the IDH wild-type is generally more invasive than the IDH mutant.

VI. Discussion

We propose a multi-modal contrastive learning framework that exploits the multi-modal features extracted from

TABLE V
Ablation Experiments

Ablation experiments	AUC	ACC	SEN	SPE
Multi-modal input				
f^I+ MLP	0.877	0.802	0.810	0.781
f^P+ MLP	0.858	0.813	0.809	0.812
f^B+ MLP	0.869	0.810	0.821	0.781
$f^I + f^P$+ MLP	0.874	0.828	0.833	0.812
$f^B + f^P$+ MLP	0.871	0.819	0.833	0.781
$f^I + f^B$+ MLP	0.876	0.836	0.860	0.750
$f^I + f^P$+ MLP+ L_{contr}	0.929	0.853	0.893	0.750
$f^B + f^P$+ MLP+ L_{contr}	0.887	0.836	0.869	0.688
$f^I + f^B$+ MLP+ L_{contr}	0.894	0.845	0.833	0.875
Multi-modal learning				
Multi-modal (no L_{multi}) + MLP	0.910	0.836	0.833	0.844
Multi-modal (no attention) + MLP	0.924	0.853	0.869	0.813
Population graph classification				
Multi-modal (full model) + MLP	0.936	0.862	0.893	0.781
Manual correction of tumor masks				
Proposed framework (auto-mask)	0.919	0.888	0.893	0.875
Fulll framework				
Proposed framework	**0.962**	**0.905**	**0.917**	**0.875**

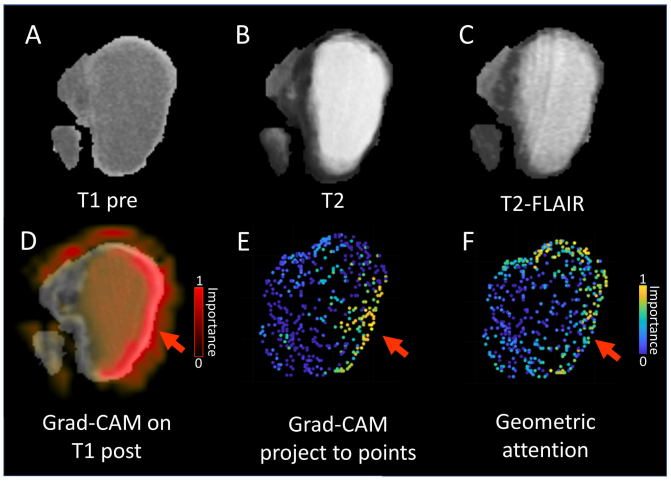

Fig. 7. An case example of interpretation of image and geometric encoders. A. pre-contrast T1; B. T2; C. T2-FLAIR. D. The Grad-CAM heatmap overlaid on post-contrast T1. E. The Grad-CAM voxels projected to the points cloud. E. Points attention generated by the geometric encoder.

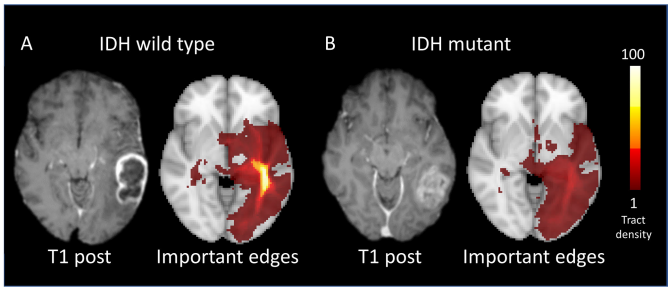

Fig. 8. Examples of IDH mutant and wild-type. A IDH wild-type. B. IDH mutant. Voxel distribution of disrupted tracts with over 50% probability of importance are indicated.

the tumor image, points cloud and global brain networks for predicting glioma genotype. We firstly develop a novel self-supervised learning approach to construct brain networks from anatomical multi-sequence MRI. Moreover, tumor-related brain network features are extracted by developing hierarchical graph attention for the brain network encoders. Further, we design a bi-level multi-modal contrastive loss that could align tumor-related network features with focal

tumor features across the domain gap. Finally, we construct a population graph integrating the multi-modal features and predict patients' genotype. Our learning framework achieves the highest performance compared to other state-of-the-art methods and benchmark models.

Previous studies show that tumor geometric features demonstrate crucial value in characterizing tumors. In our experiments, although a single geometric encoder does not perform the best, combining image and geometric encoders shows high performance, which suggests the benefit of including geometric data to extract relevant features. The interpretative visualization shown in Fig.7 indicates the agreement between points cloud and image features, which could further validate the effectiveness of the multi-modal contrastive learning in aligning multiple data domains. Biologically, this could be interpreted as the association between tumor content and tumor boundaries, indicating tumor aggressiveness and invading patterns.

Brain network provides additional value to focal tumor features, as glioma is characterized by diffuse invasion. However, glioma patients frequently demonstrate concomitant pathology beyond the lesion, which challenges extracting tumor-specific features from brain networks. Different from traditional cross-modal attention [8], we design a hierarchical attention which helps transfer the geometric attention of points cloud to crossing edges and minimize the domain gap between focal tumor and brain networks. This neuroscience-inspired attention module demonstrates significance in enhancing the model performance and interpretability, shown by the ablation experiments and visualization. The hierarchical graph attention demonstrates to indicate tumor invasion across tumor boundaries (Fig. 7), which is also associated with the white matter tracts (Fig. 8).

Instead of directly applying cross-modal contrastive loss between three modalities in the same latent space, we develop a bi-level contrastive loss, tailored to perform contrastive learning at tumor and brain levels, reflecting the gradient invasion pattern. Our experiments show that our multi-modal contrastive learning outperforms the CNN-based benchmarks, validating the usefulness of properly incorporating tumor geometrics and brain networks in predicting glioma genotype.

The population graph integrates the multi-modal features. The brain network features demonstrate as the best features describing patient similarity, the importance of network features. In contrast, focal tumor features show weaker performance in characterizing patient similarity, which might be due to the remarkable tumor heterogeneity and the limited information compared to the global brain, further supporting the value of incorporating comprehensive features in the prediction.

The proposed methods have the potential for automated, rapid diagnosis and prognosis in glioma patients based on pre-treatment MRI, essential for patient risk stratification and treatment planning towards precision medicine. Further, our hierarchical graph attention could help reveal the tumor-related disruption beyond the lesion, which could help enhance more precise planning of surgery and radiotherapy, as recent studies show that identifying disrupted white matter tracts could help reveal invisible tumor invasion on the conventional MRI and indicate recurrence location [7].

This study has limitations. Firstly, due to the rarity of glioma, our training sample is smaller than other cancers, although our cohort is one of the largest in glioma. Secondly, our cohort is slightly more imbalanced (∼25% IDH mutant) than reported incidence (∼40%) [45]. We use both AUC and accuracy in evaluating our model performance, which shows comparable results, implying the model robustness. Thirdly, constructing brain networks relies on the neuroanatomy atlases, where we use an atlas with 90 brain regions, due to the limitation of computational costs. Adopting the atlases with higher resolution could further increase the framework's performance. Our future work will involve larger datasets and transfer learning to further enhance the performance. In addition, manifold or mesh-based geometric encoder could be utilized to capture features from a more detailed geometric data format.

VII. Conclusion

We present a novel multi-modal learning framework for predicting glioma genotype. Our technical contribution include: a self-supervised approach for generating brain networks from anatomical MRI; a specialized hierarchical attention module that attends to tumor related edges and nodes; a bi-level contrastive loss for minimizing the domain gap between different modalities; a weighted population graph for feature integration and patient classification. Our framework outperforms CNN-based benchmarks and published state-of-the-art models. In future, we will further develop our model to include clinical variables into the prediction.

References

[1] C. Li et al., "Intratumoral heterogeneity of glioblastoma infiltration revealed by joint histogram analysis of diffusion tensor imaging," *Neurosurgery*, vol. 85, no. 4, pp. 524–534, 2019.

[2] C. Li et al., "Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals," *Neoplasia*, vol. 21, no. 5, pp. 442–449, May 2019.

[3] C. Li et al., "Low perfusion compartments in glioblastoma quantified by advanced magnetic resonance imaging and correlated with patient survival," *Radiotherapy Oncol.*, vol. 134, pp. 17–24, May 2019.

[4] D. N. Louis et al., "The 2016 World Health Organization classification of tumors of the central nervous system: A summary," *Acta Neuropathol.*, vol. 131, no. 6, pp. 803–820, Jun. 2016.

[5] S. Liang et al., "Multimodal 3D DenseNet for IDH genotype prediction in gliomas," *Genes*, vol. 9, no. 8, p. 382, Jul. 2018.

[6] Y. Liu et al., "Altered rich-club organization and regional topology are associated with cognitive decline in patients with frontal and temporal gliomas," *Frontiers Human Neurosci.*, vol. 14, p. 23, Feb. 2020.

[7] Y. Wei et al., "Structural connectome quantifies tumour invasion and predicts survival in glioblastoma patients," *Brain*, Oct. 2022, Art. no. awac360, doi: 10.1093/brain/awac360.

[8] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 201–216.

[9] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," 2020, *arXiv:2010.00747*.

[10] A. P. Bhandari, R. Liong, J. Koppen, S. V. Murthy, and A. Lasocki, "Noninvasive determination of *IDH* and 1p19q status of lower-grade gliomas using MRI radiomics: A systematic review," *Amer. J. Neuroradiol.*, vol. 42, no. 1, pp. 94–101, Jan. 2021.

[11] G. A. Gihr et al., "Histogram analysis of diffusion weighted imaging in low-grade gliomas: In vivo characterization of tumor architecture and corresponding neuropathology," *Frontiers Oncol.*, vol. 10, p. 206, Feb. 2020.

[12] K. Chang et al., "Residual convolutional neural network for the determination of IDH status in low-and high-grade gliomas from MR imagingneural network for determination of IDH status in gliomas," *Clin. Cancer Res.*, vol. 24, no. 5, pp. 1073–1081, 2018.

[13] A. Ahmad et al., "Predictive and discriminative localization of IDH genotype in high grade gliomas using deep convolutional neural nets," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 372–375.

[14] Y. S. Choi et al., "Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics," *Neuro-Oncol.*, vol. 23, no. 2, pp. 304–313, Feb. 2021.

[15] E. T. Bullmore and D. S. Bassett, "Brain graphs: Graphical models of the human brain connectome," *Annu. Rev. Clin. Psychol.*, vol. 7, no. 1, pp. 113–140, Apr. 2011.

[16] Y. Wei, C. Li, and S. J. Price, "Quantifying structural connectivity in brain tumor patients," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 519–529.

[17] B. R. J. Van Dijken et al., "Subventricular zone involvement characterized by diffusion tensor imaging in glioblastoma," *World Neurosurgery*, vol. 105, pp. 697–701, Sep. 2017.

[18] X. Gu, H. Knutsson, M. Nilsson, and A. Eklund, "Generating diffusion MRI scalar maps from T1 weighted images using generative adversarial networks," in *Proc. Scandin. Conf. Image Anal.* Cham, Switzerland: Springer, 2019, pp. 489–498.

[19] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[20] J. Ma, X. Zhu, D. Yang, J. Chen, and G. Wu, "Attention-guided deep graph neural network for longitudinal Alzheimer's disease analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 387–396.

[21] C. Cabanes et al., "The CORA dataset: Validation and diagnostics of in-situ ocean temperature and salinity measurements," *Ocean Sci.*, vol. 9, no. 1, pp. 1–18, Jan. 2013.

[22] S. Parisot et al., "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer's disease," *Med. Image Anal.*, vol. 48, pp. 117–130, Aug. 2018.

[23] Y. Wei, Y. Li, X. Chen, C.-B. Schönlieb, C. Li, and S. J. Price, "Predicting isocitrate dehydrogenase mutation status in glioma using structural brain networks and graph neural networks," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2022, pp. 140–150.

[24] Y. Wei, C. Li, X. Chen, C.-B. Schöinlieb, and S. J. Price, "Collaborative learning of images and geometrics for predicting isocitrate dehydrogenase status of glioma," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–4.

[25] N. Tzourio-Mazoyer et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.

[26] N. Pedano et al., "TCGA-LGG dataset, The cancer imaging archive," 2016, doi: 10.7937/K9/TCIA.2016.L4LTD3TK.

[27] L. Scarpace et al., "TCGA-GBM dataset, The cancer imaging archive," doi: 10.7937/K9/TCIA.2016.RNYFUYE9.

[28] N. Shah, X. Feng, M. Lankerovich, R. B. Puchalski, and B. Keogh, "Ivy GAP dataset. The cancer imaging archive," doi: 10.7937/K9/TCIA.2016.XLWAN6NL.

[29] 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, pp. 1–13, Sep. 2017.

[30] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[31] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[32] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.

[33] B. B. Avants, N. Tustison, and G. Song, "Advanced normalization tools (ANTs)," *Insight J.*, vol. 2, pp. 1–35, Jun. 2009.

[34] P. Prasanna et al., "Mass effect deformation heterogeneity (MEDH) on gadolinium-contrast T1-weighted MRI is associated with decreased survival in patients with right cerebral hemisphere glioblastoma: A feasibility study," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Feb. 2019.

[35] 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. 2004.

[36] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2021.

[37] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch geometric," 2019, *arXiv:1903.02428*.

[38] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[39] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[41] H. Peng et al., "Predicting isocitrate dehydrogenase (IDH) mutation status in gliomas using multiparameter MRI radiomics features," *J. Magn. Reson. Imag.*, vol. 53, no. 5, pp. 1399–1407, 2021.

[42] G. Lin, Z. Zheng, L. Chen, T. Qin, and J. Song, "Multi-modal 3D shape clustering with dual contrastive learning," *Appl. Sci.*, vol. 12, no. 15, p. 7384, Jul. 2022.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[44] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, p. 9240.

[45] Y.-Q. Liu et al., "Gene expression profiling stratifies IDH-wildtype glioblastoma with distinct prognoses," *Frontiers Oncol.*, vol. 9, p. 1433, Dec. 2019.