Mutual Contrastive Low-rank Learning to Disentangle Whole Slide Image Representations for Glioma Grading

Lipei Zhang[1]
lz452@cam.ac.uk
Yiran Wei[2]
yw500@cam.ac.uk
Ying Fu[3]
fuying@bit.edu.cn
Stephen Price[2]
sjp58@cam.ac.uk
Carola-Bibiane Schönlieb[1]
cbs31@cam.ac.uk
Chao Li (corresponding author)[1,2]
cl647@cam.ac.uk

[1] Department of Applied Mathematics and Theoretical Physics
University of Cambridge
[2] Department of Clinical Neuroscience
University of Cambridge
[3] School of Computer Science and Technology
Beijing Institute of Technology

Abstract

Whole slide images (WSI) provide valuable phenotypic information for histological assessment and malignancy grading of tumors. The WSI-based grading promises to provide rapid diagnostic support and facilitate digital health. Currently, the most commonly used WSIs are derived from formalin-fixed paraffin-embedded (FFPE) and Frozen section. The majority of automatic tumor grading models are developed based on FFPE sections, which could be affected by the artifacts introduced from tissue processing. The frozen section exists problems such as low quality that might influence training within single modality as well. To overcome these problems in the single modal training and achieve better multi-modal and discriminative representation disentanglement in brain tumor, we propose a mutual contrastive low-rank learning (MCL) scheme to integrate FFPE and frozen sections for glioma grading. We first design a mutual learning scheme to jointly optimize the model training based on FFPE and frozen sections. In this proposed scheme, we design a normalized modality contrastive loss (NMC-loss), which could promote to disentangle multi-modality complementary representation of FFPE and frozen sections from the same patient. To reduce intra-class variance, and increase inter-class margin at intra- and inter-patient levels, we conduct a low-rank (LR) loss. Our experiments show that the proposed scheme achieves better performance than the model trained based on each single modality or mixed modalities without reducing the efficiency of inference, and even improves the feature extraction in classical attention-based multiple instances learning methods (MIL). The combination of NMC-loss and low-rank loss outperforms other typical contrastive loss functions. The source code is in https://github.com/uceclz0/MCL_glioma_grading.

1 Introduction

Glioma is the most frequent malignant primary brain tumor, characterized by remarkable infiltration and tumor heterogeneity [18, 19, 20]. According to the World Health Organization (WHO) classification, glioma is classified into four grades, where grade IV represents the most aggressive type, and lower grade glioma (LGG), i.e., grades II and III, are less aggressive [18, 19]. Glioma grading has crucial significance for treatment planning and risk stratification towards precision medicine [15, 16, 37]. The current practice of glioma grading is based on the histology assessment of tumor specimens, which is time-consuming and requires high professional expertise. Therefore, an accurate and automatic approach to glioma grading based on the WSIs promises to provide rapid diagnostic support for timely clinical decision-making. Furthermore, a computer-assisted method could help facilitate digital health and enhance the accessibility of medical resources.

The FFPE tissue section is generally used as the diagnostic standard in clinical practice, which can be used for long-term storage because of formalin-fixed paraffin-embedded. Due to the gigapixel size of the images and the complexity of tumor tissue, splitting WSI into small patches is used as an effective method in model training [13, 33]. By using this operation, previous studies proposed machine learning approaches based on feature engineering [25, 34]. Although providing reasonable performance, these approaches were limited to model generalizability due to the less robust features extracted from diverse tumor tissue. Recently, most state-of-the-art models employed the deep transfer learning approach to transfer the pre-trained weights from ImageNet [8, 13, 27, 33]. These deep learning methods usually neglect the correlation among different instances because patches are typically described by weak annotation such as a global label. Therefore, some researchers proposed some attention-based multiple instance learning (MIL) methods by integrating all patches from one WSI [12, 21, 31]. However, all previous studies only consider extracting features or classification on FFPE sections. These studies might be affected by the bias of FFPE tissue. Moreover, a single section modality may not facilitate learning relevant image representations for tumor grading. Importantly, the artifacts introduced by formalin could affect the interpretation of histological specimens [22]. Specifically, the procedure of prefixation and fixation could influence the morphological quality of FFPE specimens [3]. These artifacts could pose particular challenges to the model training based on WSI. In parallel, another modality of WSI from frozen tissue procedure provides a rapid approach to tumor grading to guide intra- or peri-operative clinical decisions. Although frozen sections typically contain limited tissue, the sample hydration and cellular morphology of the frozen tissue can be preserved at a natural state [7], which may be crucial for tumor grading. However, the frozen sections also exist that some potential problems, such as poor quality, influence the performance [20]. Hence, adopting two modalities could promote extracting significantly complementary information for tumor grading by a multi-modality training scheme.

The rising mutual learning scheme promises to jointly optimise WSI representations across FFPE and frozen sections. Mutual learning [41] was initially proposed to facilitate direct knowledge distillation based on joint training scheme. Later, a hierarchical architecture with multiple classifier heads was proposed to improve model generalisation [32]. Additionally, peer mutual learning was proposed for online unified knowledge distillation [36]. Mutual learning was also successful in classification tasks based on audio-video data [2]. However, the classic losses used in mutual learning may not be able to extract extra complementary information from the latent spaces of multi-modality. On the other hand, contrastive loss [5, 11] has shown the capacity of extracting complementary representations

from latent vectors. Therefore, introducing a contrastive loss into mutual learning schemes could allow learning complementary representations jointly from multiple modalities with better generalisation, such as joint learning on visual-textual information [39] and predicting isocitrate dehydrogenase mutation of glioma [35]. Here we hypothesise that a mutual contrastive learning scheme could achieve better performance in tumour grading based on FFPE and frozen sections.

However, the contrastive loss is only about the variance of inter-modality rather than intra-class. Based on the diversity characteristic of brain tumors, cancer cells are genetically aberrant and can be divided into different sub-types at each grade [26], and there are relatively similarities between different grades [23], which causes intra-class variance and decreases the inter-class margin. There are potential solutions that aim at solving these problems, such as pairwise or triplet losses [30]. As a result, they carry an extra computation and learning burden in selecting and computing multiple pairs or triplets. In addition, the noised images will introduce uncertainty in the training phase as well. Therefore, some researchers adapted low-rank constraint to explicitly improve the discriminative capacity on natural images without specific pairing and reduce uncertainty from noised data. The low-rank constraint is achieved via a linear transformation enforcing the minimum rank of each class feature sub-matrix, and an orthogonalization constraint on the matrix of features of all classes [14, 28]. Meanwhile, unlike the well-known cross-entropy loss computed on each paired data vectors individually, low rank is able to globally optimize the lowest-rank representation on a collection of vectors, which will be more robust for noised data [17]. Based on these methods, a multi-modality mutual contrastive low-rank learning scheme becomes promising, which could simultaneously achieve disentangling representation from multi-modal and reducing intra-class variance, increase inter-class margin at intra- and inter-patient levels.

In our paper, we first hypothesize that integrating FFPE and frozen sections could train a more robust model to learn the high-level representations reflecting tumor malignancy, with less bias from the artefacts caused by tissue processing and low quality. To achieve this goal, we design a parallel mutual learning scheme to facilitate the integration of FFPE and frozen sections in model training. In this scheme, we design a normalised multi-modal contrastive loss (NMC-loss) to disentangle representation with the sphere projection [10]. Meanwhile, we adopt low-rank loss to promote latent vectors from the same class to lie in a linear sub-space by lowering the matrix rank and latent vectors at inter-class being in an orthogonal sub-space, which achieves better discriminative representation disentanglement at the intra- and inter-patient levels. Moreover, our model could perform prediction based on single modality in the testing phase without increasing the inference time and parameter number of the backbone, promising to tackle the challenge of data scarcity.

To our best knowledge, this is the first multi-modality mutual contrastive learning approach for glioma grading in the field of digital pathology. Our contributions include:

- a mutual contrastive low-rank learning (MCL) scheme for joint optimization of model training based on the WSIs of FFPE and frozen sections.

- an NMC loss to improve the ability to disentangle multi-modality representations in the mutual learning process.

- a low-rank loss to reduce intra-class variance, and increase inter-class margin at intra- and inter-patient levels.

2 Methods

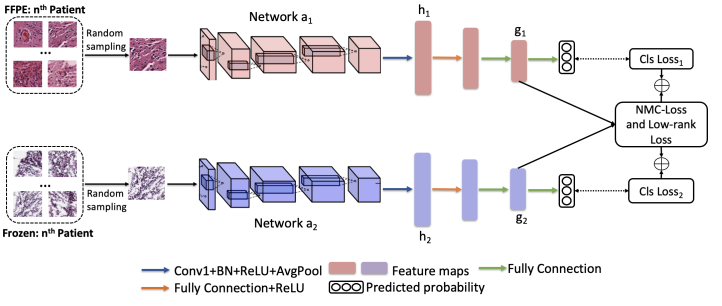

Figure 1: The pipeline of gliomas grading on FFPE sections and frozen sections. The FFPE and frozen patches will be randomly sampled from the bags from the same patient. Two sorts of patches are input into two networks without weight-sharing. The feature vectors from different layers play different roles in the representation disentanglement and classification.

The proposed parallel mutual optimization learning scheme is shown in Fig. 1. Initially, a random sampling strategy is adopted for the mutual training. The FFPE and frozen patches will be randomly sampled from the bags of the same patient. The paired images are input into the Network α_1 and α_2, respectively. In each branch, the set of feature vectors (h_1 and h_2) after the first fully connected layer and ReLU are input into the next two fully connected layers respectively for multi-modal representation disentanglement and low-rank optimization. Combining contrastive loss, low-rank loss and classification loss could optimize each modality network jointly via the back-propagation. We introduce details of the main design in the following parts.

Non-linear representation. We adopt three fully connections with ReLU in our scheme to promote better non-linear projection of latent vectors for representation learning. In each modal network, the latent vectors after average pooling are transformed by first fully connection, ReLU and second hidden layer with the same dimensional transformation, so that we can obtain a representative vector with the non-linear projection ($z_i = g(h_i) = W^2\sigma(W^1 h_i)$), where σ is s a ReLU non-linearity. Referred by the experiments of SimCLR [5], a non-linear operator can remove variant information, e.g., the color or orientation of objects resulting from various staining procedures from multiple centers. Therefore, these latent vectors can be more suitable for representation disentanglement and the final classification head can leverage the nonlinear transformation to maintain more useful information in h, which could boost the performance.

NMC-loss. For this design, formally given a mini-batch of size N, we firstly view the FFPE $x_1^a,...,x_K^a$ and the frozen section , $x_1^b,...,x_K^b$ images as images sampled from different augmented views on the same patient. There are $2N-1$ pairs totally, among which we can regard the corresponding augmented sample x_i^b as a positive pair x_k^a, x_k^b and other $2N-2$ pairs are negative samples. In standard contrastive loss definition presented in [5, 11], $l2$ normalization was used for reference and augmented images. However, the image pairs used in self-supervised learning keep unified features and distribution, and $l2$ normalization can scale latent vector into a valid range. In our task, the image pairs include two domains with different distributions. Therefore, we adopt layer normalization, as shown in Eq. 1, to re-centre and rescale the latent space into the same sphere, which can improve the efficiency of

disentanglement.

$$\hat{g}_n = \frac{g_n - \mu_n}{\sqrt{(\sigma_n)^2 + \varepsilon}} \tag{1}$$

where $n \in \{1,2\}$ and g_n denotes the latent vector from FFPE or frozen section. μ and σ are the mean and variance of each batch. Moreover, we denote $sim(\hat{g}_1, \hat{g}_2) = \frac{\hat{g}_1^T \hat{g}_2}{||\hat{g}_1|| \cdot ||\hat{g}_2||}$ as the cosine similarity between $\hat{g}_1$ and $\hat{g}_2$. The NMC-loss function can be defined as:

$$L_k^a = -log \frac{exp(sim(\hat{g}_k^a, \hat{g}_k^b)/\tau)}{\sum_{i \in I} \mathbb{1}_{i \neq k} exp(sim(\hat{g}_k^a, \hat{g}_i^b)/\tau)} \tag{2}$$

where $\mathbb{1}_{i \neq k} \in \{0,1\}$ is an indicator, which values $\mathbb{1}$ only when $i \neq k$. We also define τ as a temperature hyper-parameter. To identify all positive pairs in this batch, the NMC-loss is further defined as (L_k^b follows the same calculation with L_k^a) :

$$L_{nmc} = \frac{1}{2N} \sum_{k=1}^{K} (L_k^a + L_k^b) \tag{3}$$

Low-rank loss. Inspired by the idea of low rank of representations [14, 17], we can consider that successful training of the network would result in the classifier vectors remaining orthogonal at the end if non-linear representation g_n can be in the orthant and low-rank. More precisely, we can consider a feature embedding of each modality $X^a = [x_1^a | x_2^a | ... | x_N^a]$, where each column $x_i^a \in R^d, i = 1, ..., N$, $a \in [1,2]$ and $|$ represents vertical concatenation. The X^a is obtained from a given training samples Y with minibatch size N, $X = \phi(Y;\theta)$, and X is the $N \times D$ deep embedding from extractor ϕ with parameter θ. We further assume that X_c^a, Y_c^a are the sub-feature matrices and input 5data respectively belonging to grade c and modality a. To achieve better discriminative representation disentanglement in intra- and inter-patient levels, the sub-matrices from two modalities will be concatenated from a vertical direction such as $M = [X^1 | X^2]$. In each minibatch, our low-rank loss can be defined as the following equations:

$$\begin{aligned} L_{lr} &= \sum_{c=1}^{C} max(\Delta, ||M_c||_*) - ||M||_* \\ &= \sum_{c=1}^{C} max(\Delta, ||[\phi(Y_c^1;\theta) | \phi(Y_c^2;\theta)]||_*) - ||[\phi(Y^1;\theta) | \phi(Y^2;\theta)]||_* \end{aligned} \tag{4}$$

Where $||.||_*$ means the matrix nuclear norm (the sum of the singular values) $\Delta \in R$ denotes a bound on the intra-class nuclear loss that can avoid the training collapse resulting from feature value to zero. In our experiments, we set $\Delta = 1$. The first term in loss function can minimize the rank of each grade feature subspace and the second term promotes inter-class to be linearly orthogonal.

To further clarify the optimization in backpropagation, we can calculate a simplified subgradient of the nuclear norm. Based on the SVD decomposition and deduction from [14], the descent direction can be defined as the following equation:

$$gL_{lr}(M) = \sum_{c=1}^{C} [Z_c^{(l)} | U_{c1} V_{c1}^T | Z_c^{(r)}] - U_1 V_1^T \tag{5}$$

Where, $Z_c^{(l)}$ and $Z_c^{(r)}$ denote fill matrices of zeros to keep the original dimension of M. U_1 and V_1 are the principal left and right singular vectors from M and U_{c1} and V_{c1} are chosen by the singular value being greater than a fixed threshold.

Model optimization. Our learning scheme consists of two functions to achieve joint optimization in two classification tasks. Each function consists of a cross-entropy (CE) loss with a Taylor Softmax [9], and an NMC-loss and a low rank loss. The Taylor Softmax CE can smooth labels to reduce over-fitting and it already has been proved effectiveness in many competition solutions. This loss can be formed as Eq. 3 and the total loss function of each modality can be expressed as Eq. 4:

$$L_{cls}(f(x),y) = \sum_{i=1}^{t} \frac{(1-f_y(x))^i}{i} \tag{6}$$

$$\begin{aligned} L_{FFPE} &= L_{cls_1} + L_{nmc} + L_{lr} \\ L_{Frozen} &= L_{cls_2} + L_{nmc} + L_{lr} \end{aligned} \tag{7}$$

where the $f_y(x)$ denotes the y-th element of $f(x)$ and $f(.)$ is a CNN with the classification layer. t is the term number of the Taylor series and we set $t = 3$ as the same as the setting from the original paper.

3 Experimental Setup

3.1 Datasets

We utilized the WSI of glioblastoma (GBM) and LGG from the Cancer Genome Atlas (TCGA) dataset [1], with clinical details and Hematoxylin and Eosin (HE) stained sections available. We only selected 499 patients (108 grade II, 94 grade III, and 297 grade IV) with both FFPE and frozen sections available.

For data pre-processing, we designed three steps: 1) transforming a low-dimension version of WSI into HSV color space and separating HE-stained tissue from the background using Otsu's Binarization on the saturation channel [4]; 2) patching a number of non-overlapping 500×500 instance-level images at $20\times$ magnification; 3) a blob detection procedure [33] to further remove redundant patches containing insufficient tissue. The numbers of finally included patches were 1,680,714 for FFPE sections and 483,886 for frozen sections.

To evaluate the proposed scheme, the dataset was randomly divided into 319 patients for the training set, 80 patients for the validation set and 100 patients for the testing set. In testing set, it included 27 grade II, 25 grade III, and 48 grade IV. Moreover, to increase sample size, we cropped sub-regions of patches into a size of 224×224. In addition, we applied data augmentation techniques (random rotation of 90°, 180°, 270°, random flipping image along axis, shift hue saturation value and brightness contrast) to increase the training sample size.

3.2 Training Details

The training environment was based on PyTorch 1.6.0 backend with acceleration by Nvidia RTX 3090. The batch size was set to 32, corresponding to 32 pairs of FFPE and frozen images in the mutual training, while 32 of FFPE or frozen images in single training and mixed

training. The input was $224 \times 224 \times 3$. We used the loss function described in. Eq.4 for our experiments. The number of training epochs was 10, and the optimizer was Adam with default parameters. Cosine annealing warm restarts were adopted with an initial learning rate of 1.6×10^{-4}.

We trained different CNN backbones with single-modal input training (baseline) [13, 33], mixed-modal training and our proposed mutual training scheme. Meanwhile, some comparisons with state of art methods (SOTAs), such as attention MIL (A-MIL) [12], TransMIL [31] and CLAM [21], were introduced. In the original papers, the feature vector of each patch was from the backbone with ImageNet pre-trained weights. Therefore, we further used the trained weights from the different learning schemes to extract feature vectors at patch-level. Moreover, we compared the performance of different metrics loss in mutual training such as Kullback-Leibler divergence [41], marginal triplet loss [29], NT-logistic loss, NT-Xent loss [5] and angular margin contrastive loss (AMC) [6] as well.

3.3 Testing Details

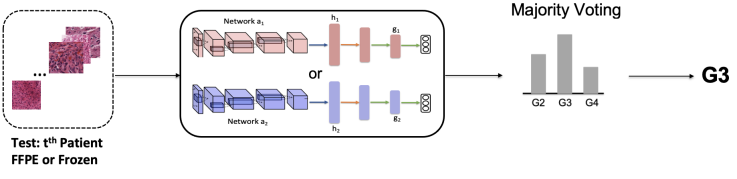

Figure 2: The pipeline of gliomas grading on FFPE section and frozen section.

The details of the inference phase are shown in Fig. 2. FFPE and frozen sections were classified separately by Network a_1 or a_2 with respective classification heads. After predicting all images in a patient's bag, the number of predicted grades was counted by the histogram and majority voting was used to determine the final predicted tumor grade. Moreover, in the comparison study, we also followed protocols in A-MIL [12], CLAM [21] and TransMIL [31] to train and test at patient-level feature matrices.

4 Results

4.1 Quantitative Results

For comparison, the evaluation metrics on the CNN backbone with single input training, mixed training, our proposed scheme, SOTAs and combinations of different learning schemes and SOTAs are shown in Table 1. We chose EfficientNet-B0 to evaluate the model performance in these experiments. We observe that our proposed learning scheme outperforms the single and mixed training on the given backbone. These results suggest that the performance of the single training could be limited by the information from each single specific modality, while the mixed training may not efficiently obtain complementary information from the batch-size learning. After being applied to the different learning schemes on SOTAs, our proposed scheme demonstrates the capability of disentangling mutual information from multi-modality and reducing intra-class variance in the latent vector space, which could increase the SOTAs performance at the instance and patient-level.

Table 1: Comparison with different learning schemes. Single training only trained and tested on the same single modality. Mixed training mean combining modalities for training on one model and separately testing in each. Mutual training followed proposed method to train and test. The ImageNet pre-trained weights, single trained weights, mixed trained weights and mutual trained weights were used in feature extraction and combined with A-MIL, TransMIL and CLAM to train and test in patient-level.

	FFPE			Frozen		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Single training	0.71	0.72	0.71	0.70	0.71	0.70
Mixed training	0.70	0.72	0.71	0.68	0.67	0.68
MCL	**0.76**	**0.77**	**0.76**	**0.74**	**0.73**	**0.74**
ImageNet + A-MIL	0.72	0.72	0.72	0.70	0.68	0.70
ImageNet + TransMIL	0.71	0.72	0.71	0.74	0.73	0.74
ImageNet + CLAM	0.71	0.72	0.71	0.69	0.70	0.69
single training+A-MIL	0.74	0.75	0.74	0.74	0.73	0.74
single training+TransMIL	0.69	0.70	0.73	0.73	0.72	0.73
single training+CLAM	0.75	0.76	0.75	0.73	0.72	0.73
Mixed training+A-MIL	0.73	0.73	0.73	0.70	0.69	0.70
Mixed training+TransMIL	0.70	0.70	0.70	0.72	0.73	0.72
Mixed training+CLAM	0.75	0.75	0.75	0.71	0.69	0.71
MCL+A-MIL	0.78	0.79	0.78	0.75	0.74	0.75
MCL+TransMIL	0.77	0.77	0.77	0.74	0.75	0.74
MCL+CLAM	**0.79**	**0.80**	**0.79**	**0.75**	**0.75**	**0.75**

To further demonstrate that our proposed NMC loss and low-rank loss could fulfill a better representation disentanglement at intra- and inter-patient levels, we compared it with other contrastive loss functions as shown in Table. 2. We observe that the NMC loss provides superior performance, benefiting the representation disentanglement. In comparison, KL-loss fails to consider the distance between positive and negative samples, which might lead to worse performance than other loss functions. As for the marginal triplet, NT-logistic loss and AMC loss, they are measured using the absolute similarity of the positive and negative samples. Although using the relative similarity may help the network optimize the balance between separating the samples of different classes in NT-Xent loss, $l2$ normalization on a single modality can not match differences between modalities. Therefore, the results of NMC-loss illustrate the advantage of layer normalization. Moreover, the low-rank loss shows capacity of the discriminative information disentanglement at intra- and inter-patient levels, compared with other contrastive loss. The combination of the NMC-loss and LR loss achieves best performance. From Table.3, a temperature (τ) of 0.5 performed the best. The models in different temperature hyper-parameter have stable accuracy and show the consistent robustness in this task. In our experiments, we found that the training will collapse if temperature is smaller than 0.05.

4.2 Visualization

To understand how our proposed scheme leverages representation disentangling in predicting tumor grades, we visualized the latent vector from single training, mixed training and

Table 2: Comparison with different contrastive loss functions

	FFPE			Frozen		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KL-loss	0.71	0.70	0.68	0.71	0.68	0.69
Marginal triplet loss	0.74	0.74	0.74	0.72	0.72	0.72
NT-Logistic loss	0.74	0.74	0.72	0.71	0.68	0.70
NT-Xent loss	0.71	0.71	0.71	0.73	0.73	0.73
AMC loss	0.75	0.76	0.75	0.70	0.70	0.70
LR loss	0.75	0.77	0.75	0.72	0.73	0.72
NMC-loss	0.75	0.75	0.75	0.72	0.71	0.72
NMC-loss + LR Loss	**0.76**	**0.77**	**0.76**	**0.74**	**0.73**	**0.74**

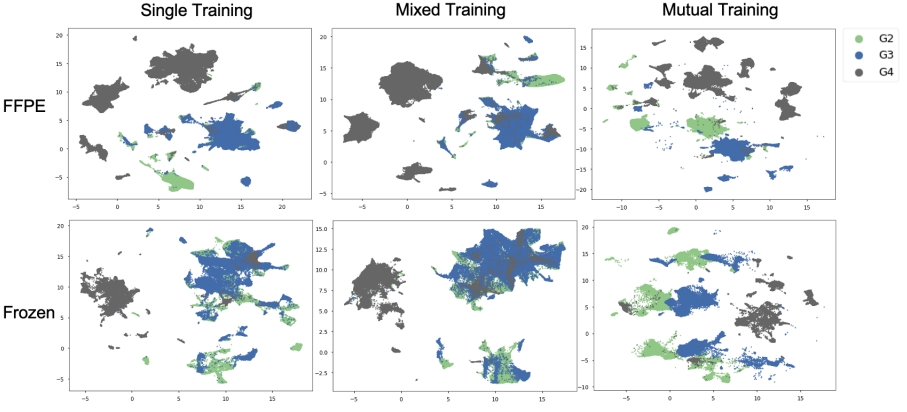

Figure 3: Mutual learning scheme interpretability in brain tumor grading. Each point is from reducing the dimension of the latent vector by UMAP method.

our proposed scheme on FFPE and frozen section images, using the Uniform Manifold Approximation and Projection (UMAP) method [24]. As shown in Fig. 3, the latent vectors are obtained from the CNN extractors. The results show that our proposed scheme could promote multi-modality and discriminative representation disentanglement, which may help the latent vector on the classification head preserve more helpful information from the same tumor grade, demonstrating closer distribution in the feature space.

The qualitative performance of the randomly selected statistical features from our proposed scheme is illustrated in Fig. 4. These salient maps are generated by classification activation maps (CAMs) [42]. The examples in Fig. 4 show that the trained model by our proposed scheme can focus on the proliferation region, which helps efficiently detect tissue morphology from the WSI.

Table 3: Sensitivity (accuracy) of temperature hyper-parameter

Temperature	1	0.5	0.1	0.05
FFPE	0.75	0.76	0.75	0.72
Frozen	0.70	0.74	0.74	0.72

5 Conclusions

In this paper, we propose a mutual low-rank contrastive clustering learning scheme to improve the performance of representation disentangling on whole slide images for tumor grading. We first develop a mutual learning scheme to extract relevant image representations by

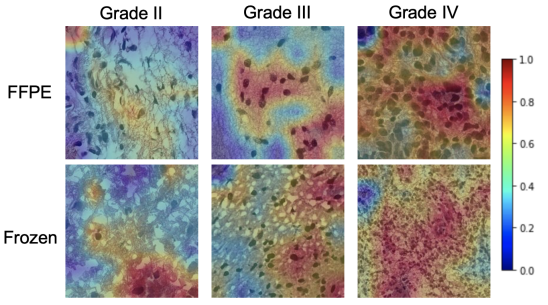

Figure 4: Image-level predicted heatmap of three tumor grades by mutual learning. Left to right: Grade II, III, and IV). The FFPE and Frozen images of each grade are from the same patients. The color bar indicates the estimated level of attention on the region.

integrating FFPE and frozen sections with complementary information. Next, We design an NMC loss, which could promote multi-modality representation disentangling within the same sphere. To further achieve discriminative representation disentanglement on the intra- and inter-patient levels, we conduct a low-rank loss. We would note that grading certain glioma are especially challenging by nature, as shown in most SOTA models, due to the tangled features across these grades in terms of morphology and cell compositions. Our proposed method has demonstrated superiority in disentangling representations compared with single or mixed training. Combined with attention-based MIL methods, our method could better extract robust features leading to better model performance. Moreover, compared with some multi-modal learning architectures that require multiple modalities in both training and testing phases, our model could perform prediction based on single modality in the testing phase, promising to tackle the challenge of data scarcity. Empirically, these loss functions could be applied to the grading of other tumours or multi-modal learning tasks in MRI and CT images. Further, as these losses can plug in at the latent vector level and do not reduce the efficiency of inference, they can be easily incorporated into other backbone networks. In the future, we will further explore mutual self-supervised architecture so that this representation learning can transfer into more downstream tasks which achieve by the attention based MIL.

References

[1] The cancer genome atlas. URL https://tcga-data.nci.nih.gov/tcga/.

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.

[3] B Paige Bass, Kelly B Engel, Sarah R Greytak, and Helen M Moore. A review of preanalytical factors affecting molecular, protein, and morphological analysis of formalin-fixed, paraffin-embedded (ffpe) tissue: how well do you know your ffpe specimen? *Archives of pathology and laboratory medicine*, 138(11):1520–1530, 2014.

[4] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds:

Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Hongjun Choi, Anirudh Som, and Pavan Turaga. Amc-loss: Angular margin contrastive loss for improved explainability in image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 838–839, 2020.

[7] Miriam Cohen, Nissi M Varki, Mark D Jankowski, and Pascal Gagneux. Using unfixed, frozen tissues to study natural mucin distribution. *JoVE (Journal of Visualized Experiments)*, (67):e3928, 2012.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2206–2212, 2021.

[10] Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Spherical knowledge distillation. 2020.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[13] Sanghyuk Im, Jonghwan Hyeon, Eunyoung Rha, Janghyeon Lee, Hojin Choi, Yuchae Jung, and Taejung Kim. Classification of diffuse glioma subtype from clinical-grade pathological images using deep transfer learning. *Sensors*, 21(10):3500, 2021.

[14] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118, 2018.

[15] Chao Li, Shuo Wang, Pan Liu, Turid Torheim, Natalie R Boonzaier, Bart RJ van Dijken, Carola-Bibiane Schönlieb, Florian Markowetz, and Stephen J Price. Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals. *Neoplasia*, 21(5):442–449, 2019.

[16] Chao Li, Jiun-Lin Yan, Turid Torheim, Mary A McLean, Natalie R Boonzaier, Jingjing Zou, Yuan Huang, Jianmin Yuan, Bart RJ van Dijken, Tomasz Matys, et al. Low perfusion compartments in glioblastoma quantified by advanced magnetic resonance imaging and correlated with patient survival. *Radiotherapy and Oncology*, 134:17–24, 2019.

[17] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Icml*, 2010.

[18] David N Louis, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvet, Bernd W Scheithauer, and Paul Kleihues. The 2007 who classification of tumours of the central nervous system. *Acta neuropathologica*, 114(2):97–109, 2007.

[19] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[20] Ming Y. Lu, Richard J. Chen, and Faisal Mahmood. Semi-supervised breast cancer histology classification using deep multiple instance learning and contrast predictive coding (Conference Presentation). In John E. Tomaszewski and Aaron D. Ward, editors, *Medical Imaging 2020: Digital Pathology*, volume 11320. International Society for Optics and Photonics, SPIE, 2020. doi: 10.1117/12.2549627. URL https://doi.org/10.1117/12.2549627.

[21] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[22] William Mathieson and Geraldine A Thomas. Why formalin-fixed, paraffin-embedded biospecimens must be used in genomic medicine: An evidence-based review and conclusion. *Journal of Histochemistry & Cytochemistry*, 68(8):543–552, 2020.

[23] Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.

[24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[25] Hojjat Seyed Mousavi, Vishal Monga, Ganesh Rao, and Arvind UK Rao. Automated discrimination of lower and higher grade gliomas based on histopathological image analysis. *Journal of pathology informatics*, 6, 2015.

[26] JG Nicholson and HA Fine. Diffuse glioma heterogeneity and its therapeutic implications. cancer discov. 2021; 11: 575–590. doi: 10.1158/2159-8290. Technical report, CD-20-1474.[Abstract][CrossRef][Google Scholar].

[27] Linmin Pei, Karra A Jones, Zeina A Shboul, James Y Chen, and Khan M Iftekharuddin. Deep neural network analysis of pathology images with integrated molecular data for enhanced glioma classification and grading. *Frontiers in oncology*, 11:2572, 2021.

[28] Qiang Qiu and Guillermo Sapiro. Learning transformations for clustering and classification. *J. Mach. Learn. Res.*, 16(1):187–225, 2015.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[31] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

[32] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[33] An Hoai Truong, Viktoriia Sharmanska, Clara Limback-Stanic, and Matthew Grech-Sollars. Optimization of deep learning methods for visualization of tumor heterogeneity and brain tumor grading through digital pathology. *Neuro-Oncology Advances*, 2 (1):vdaa110, 2020.

[34] Xiuying Wang, Dingqian Wang, Zhigang Yao, Bowen Xin, Bao Wang, Chuanjin Lan, Yejun Qin, Shangchen Xu, Dazhong He, and Yingchao Liu. Machine learning models for multiparametric glioma grading with quantitative result interpretations. *Frontiers in neuroscience*, page 1046, 2019.

[35] Yiran Wei, Chao Li, Xi Chen, Carola-Bibiane Schöinlieb, and Stephen J Price. Collaborative learning of images and geometrics for predicting isocitrate dehydrogenase status of glioma. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.

[36] Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10302–10310, 2021.

[37] Jia Wu, Chao Li, Michael Gensheimer, Sukhmani Padda, Fumi Kato, Hiroki Shirato, Yiran Wei, Carola-Bibiane Schönlieb, Stephen John Price, David Jaffray, et al. Radiological tumour classification across imaging modality and histology. *Nature Machine Intelligence*, 3(9):787–798, 2021.

[38] Yanzhe Xu, Teresa Wu, Fei Gao, Jennifer R Charlton, and Kevin M Bennett. Improved small blob detection in 3d images using jointly constrained deep learning and hessian analysis. *Scientific reports*, 10(1):1–12, 2020.

[39] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

[40] Amin Zadeh Shirazi, Mark D McDonnell, Eric Fornaciari, Narjes Sadat Bagherian, Kaitlin G Scheer, Michael S Samuel, Mahdi Yaghoobi, Rebecca J Ormsby, Santosh Poonnoose, Damon J Tumes, et al. A deep convolutional neural network for segmentation of whole-slide pathology images identifies novel tumour cell-perivascular niche interactions that are associated with poor survival in glioblastoma. *British Journal of Cancer*, 125(3):337–350, 2021.

[41] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.