Intraoperative Glioma Grading Using Neural Architecture Search and Multi-Modal Imaging

Anqi Xiao, Biluo Shen, Xiaojing Shi, Zhe Zhang, Zeyu Zhang, Jie Tian, *Fellow, IEEE*, Nan Ji, and Zhenhua Hu, *Senior Member, IEEE*

***Abstract*— Glioma grading during surgery can help clinical treatment planning and prognosis, but intraoperative pathological examination of frozen sections is limited by the long processing time and complex procedures. Near-infrared fluorescence imaging provides chances for fast and accurate real-time diagnosis. Recently, deep learning techniques have been actively explored for medical image analysis and disease diagnosis. However, issues of near-infrared fluorescence images, including small-scale, noise, and low-resolution, increase the difficulty of training a satisfying network. Multi-modal imaging can provide complementary information to boost model performance, but simultaneously designing a proper network and utilizing the information of multi-modal data is challenging. In this work, we propose a novel neural architecture search method DLS-DARTS to automatically search for network architectures to handle these issues. DLS-DARTS has two learnable stems for multi-modal low-level feature fusion and uses a modified perturbation-based derivation strategy to improve the performance on the area under the curve and accuracy. White light imaging and fluorescence imaging in the first near-infrared window (650-900 nm) and the second near-infrared window (1,000-1,700 nm) are applied to provide multi-modal information on glioma tissues. In the experiments on 1,115 surgical glioma specimens, DLS-DARTS achieved an area under the curve of 0.843 and an accuracy of 0.634, which outperformed manually designed convolutional neural networks including ResNet, PyramidNet, and EfficientNet, and a state-of-the-art neural architecture search method for multi-modal medical image classification. Our study demonstrates that DLS-DARTS has the potential to help neurosurgeons during surgery, showing high prospects in medical image analysis.**

***Index Terms*— Deep learning, glioma grading, intraoperative imaging, multi-modal imaging, neural architecture search, NIR-II fluorescence imaging.**

Manuscript received 21 January 2022; revised 29 March 2022; accepted 4 April 2022. Date of publication 11 April 2022; date of current version 30 September 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0205200; in part by the National Natural Science Foundation of China (NSFC) under Grant 62027901, Grant 81930053, Grant 92059207, and Grant 81227901; in part by the Beijing Natural Science Foundation under Grant JQ19027; in part by CAS Youth Interdisciplinary Team under Grant JCTD-2021-08; in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA16021200; in part by the Zhuhai High-Level Health Personnel Team Project under Grant Zhuhai HLHPTP201703; and in part by the Innovative Research Team of High-Level Local Universities in Shanghai. *(Corresponding authors: Jie Tian; Nan Ji; Zhenhua Hu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the study was approved by the Ethics Committee of Beijing Tiantan Hospital, Capital Medical University. This study on patients with glioma was explored in a clinical trial (ChiCTR2000029402) in China.

Anqi Xiao, Biluo Shen, Xiaojing Shi, and Zhenhua Hu are with the CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging and The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xiaoanqi2020@ia.ac.cn; shenbiluo2019@ia.ac.cn; shixiaojing2017@ia.ac.cn; zhenhua.hu@ia.ac.cn).

Zhe Zhang is with the Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China, and also with the China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Beijing 100070, China (e-mail: doctor_zneuro@126.com).

Zeyu Zhang is with the CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging and The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing 100191, China (e-mail: zhang.zey.doc@gmail.com).

Jie Tian is with the CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging and The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, also with the Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing 100190, China, and also with the Engineering Research Center of Molecular and Neuro Imaging of Ministry of Education, School of Life Science and Technology, Xidian University, Xi'an 710071, China (e-mail: tian@ieee.org).

Nan Ji is with the Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China, also with the China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Beijing 100070, China, and also with the Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing 100191, China (e-mail: jinan@bjtth.org).

Digital Object Identifier 10.1109/TMI.2022.3166129

I. Introduction

GLIOMA is the primary central nervous system tumor arising from glial or precursor cells and accounts for 70% of adult malignant primary brain tumors [1]. The five-year relative survival rate for glioblastoma, the most lethal glioma, is only 6.8% between 2012 and 2016 in the United States [2]. World Health Organization (WHO) classifies glioma into Grade I-IV according to the invasive histopathology results. Diagnosis of glioma grades facilitates clinical treatment planning and prognosis, providing benefits to patients [3], [4].

Neurological microsurgery under white light (WL) is the major treatment modality to improve patients' survival. The precise grading of gliomas during surgery can guide surgeons to determine the maximum excision area during operation [5], reducing tumor residual and early recurrence caused

by less excision and damage to the normal physiological function of the patient caused by over excision. However, it is very challenging for neurosurgeons alone to determine glioma grades directly during operation. Intraoperative pathological examination of hematoxylin and eosin (H&E) stained frozen tissue sections is reliable and widely used, but it is time-consuming (at least 20 minutes) and requires complex procedures and specialists to get pathological results [6]. Moreover, freezing tens or hundreds of samples is not practical during surgery. These issues limit the applications of pathological examination of frozen tissues for real-time intraoperative diagnosis, especially for numerous tissue samples. Therefore, a fast and precise diagnosis of tumor tissues during surgery is crucial.

Optical imaging combined with disease diagnosis and treatment becomes a very promising strategy [7]–[13]. Optical imaging methods have been used in many pre-clinical studies [14]–[19]. Recent studies suggest that near-infrared (NIR) fluorescence imaging provides images that contain different and more information than WL images, and to some extent, it helps to guide the surgeons during the surgery in real-time [20]–[22]. However, NIR fluorescence imaging for real-time glioma grading has rarely been explored and achieved.

With the rapid development of deep learning, many researchers have applied convolutional neural networks (CNNs) to analyze a wide range of medical images, including CT [23], MRI [24], ultrasound [25], and histopathological images [26], and showed encouraging results. CNNs have also been extensively studied to offer the grading [27]–[30] and genetic information [31], assist pathological diagnosis [32], [33], help determine prognosis and guide therapy [34]. However, deep learning techniques are rarely applied to intraoperative NIR imaging of glioma. This may be explained by that NIR images are harder to acquire compared with generally used intraoperative imaging methods such as ultrasound, which limits the scale of the dataset. Besides, the quality of intraoperative NIR images is severely degraded by ambient light and background noise. The resolution of intraoperative NIR images is low compared with pathology. These issues increase the difficulty of training a satisfying neural network for diagnosis. Multi-modal imaging provides more information compared with single-modal imaging, which can alleviate these issues. But designing a proper network to simultaneously handle the problems of NIR images and use the information of multi-modal data is challenging.

Neural Architecture Search (NAS) is a subarea of automated machine learning [35], which is a newly rising technique in artificial intelligence. This technique can automatically explore and evaluate a large number of networks in the search space that have never been studied before, and search for the appropriate network architectures for the target task. It has achieved better performance than manually designed CNNs in image classification [36], [37], object detection [38], and semantic segmentation [39]. Some recent studies in the medical field have also used NAS and achieved remarkable results in the segmentation of MRI and CT images [40]–[42]. This technique has a high potential to unleash the power of deep learning in brand-new scenarios or on the application of images of new imaging modalities.

Herein, we propose a NAS-based method for glioma grading intraoperatively, which we call Double-Learnable-Stem DARTS (DLS-DARTS). We also apply multi-modal imaging to offer more information for improving the performance. Current manually designed CNNs use complex structures to capture features of different modes in the multi-modal analysis [43], [44], which requires expert experiences to design concrete modules to extract features of different modes to achieve remarkable performance. By contrast, we design two learnable stems that automatically learn the operations and connections to process and combine low-level features of different modes, and further fuse them in deeper layers. WL, the first NIR fluorescence window (NIR-I, 700-900 nm), and the second NIR fluorescence window (NIR-II, 1,000-1,700 nm) glioma specimen images are collected simultaneously to construct a multi-modal image dataset. Using this dataset, DLS-DARTS can automatically search for the appropriate architectures of network stems and cells, and establish a fast and accurate grading model using the discovered architectures. The overall pipeline of our method is illustrated in Fig. 1. The main contributions of our work are summarized as follows:

- We develop a NAS method DLS-DARTS for intraoperative glioma grading. DLS-DARTS has two learnable stems to fully utilize the features of multi-modal glioma images and a modified perturbation-based derivation strategy to improve the accuracy and AUC of derived architectures. Experimental results show that DLS-DARTS achieves the highest accuracy and AUC compared with manually designed CNNs and a state-of-the-art NAS method for multi-modal medical image analysis.
- Multi-modal imaging, including WL, NIR-I, and NIR-II imaging, has been employed to provide abundant and complementary information for grading models to learn. The usage of multi-modal imaging significantly improves accuracy and AUC compared with single-modal imaging.
- DLS-DARTS shows high promise for intraoperative glioma grading with rapid and effective diagnosis. Besides, it offers real-time assistance to neurosurgeons during surgery. Our study also shows the effectiveness of NAS and its great prospects in medical imaging analysis.

II. Methodology

One of the widely used NAS methods is Differentiable ARchiTecture Search (DARTS) [36]. DARTS is based on gradient and shows remarkable results in classification tasks. It contains two phases: the search phase and the training phase. The former is to search for the appropriate architecture of the network based on the input dataset, and the latter is to train the searched network. We refer to the full network in the search phase as supernet. The main structure of the supernet of DARTS is composed of a stem for mapping input data to features, a set of cells for feature extraction, and a head for classification.

Inspired by this method, we propose Double-Learnable-Stem DARTS (DLS-DARTS), which follows the design of

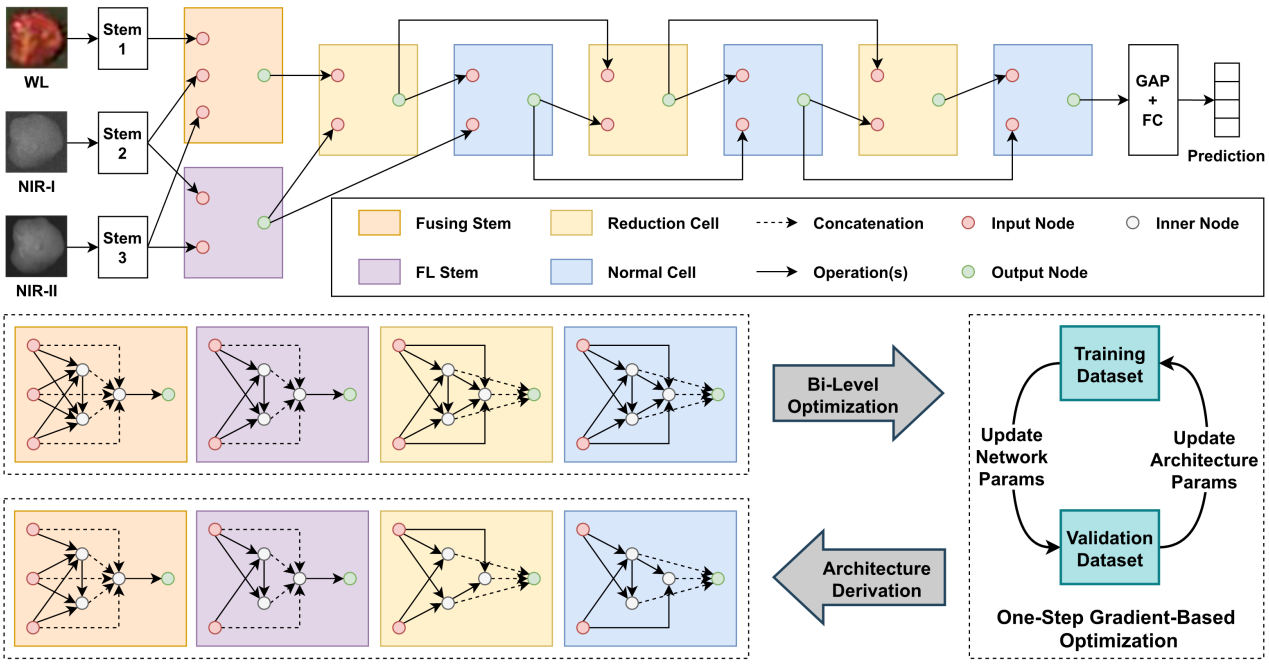


Fig. 1. An illustration of the overall pipeline of DLS-DARTS for glioma grading on multi-modal images. The total number of cells L is 6 in the supernet. Features of wight light (WL) and NIR fluorescence images are preliminarily processed by two learnable stems and further merged in the deeper cells. Fluorescence images are normalized to range 0-255 for visualization.

DARTS, while designing two learnable stems for automatically utilizing features of intraoperative multi-modal data. Besides, a modified perturbation-based architecture selection strategy is used to derive the architectures from the supernet of DLS-DARTS. It should be noted that the original DARTS only has a manually designed stem to map input data to features, which treats every mode equally and neglects the integration of different data modes when analyzing multi-modal images. However, DLS-DARTS specifically designs two learnable stems with similar architectures of cells in the network, which preliminarily merges the low-level features of different modes, and further fuses them in deeper layers to boost the performance in multi-modal image analysis.

A. Search Space of DLS-DARTS

1) Macro Architecture of Supernet: Formally, DLS-DARTS decomposes the supernet into L cells and two learnable stems. The L cells include normal cells and reduction cells, with the former mainly for calculation and the latter mainly for downsampling. Both types of cells use a weight-sharing strategy, which reduces the search overhead and helps the supernet to converge. One learnable stem is for automatically fusing features of all modes of images (WL, NIR-I and NIR-II). Since fluorescence images share similar features, we also design a learnable stem specially for extracting the features of fluorescence images. The two learnable stems are called Fusing Stem and FL Stem in the following. They both have similar architectures to the cells while having different connections that are specially designed for multi-modal low-level feature fusion.

An illustration of the architecture of DLS-DARTS with $L = 6$ is shown in Fig. 1. Different from DARTS, the cells are started with a reduction cell rather than a normal cell since the learnable stems have already extracted the low-level features from input data. The output of Fusing Stem is used as the input of the first cell, while the one of FL Stem is used as the input of the first two cells. This design allows the model to extract and preliminarily merge different low-level features of different modes, and further integrate them in deeper layers. The cells are divided into 3 groups, with one reduction cell followed by K_{search} normal cells in each group. They are used to construct the main structure of the supernet. Thus, the number of cells in the supernet can be denoted as $L = 3 \times (K_{search} + 1)$. A head with a global average pooling (GAP) and a fully connected layer (FC) follows the cells for classification.

2) Micro Architecture of Cells: Due to the learnable stems having similar architectures of cells, we first introduce the micro architectures of cells. In general, each cell contains N nodes to construct a directed acyclic graph (DAG), and each node denotes some features. Define a set of pre-defined operations O, where each operation $o \in O$ is performed on features. In the DAG, each edge that connects two nodes represents a set of operations in O. Each operation o is given a hyper-parameter α^o to control the weight. The nodes, except the first two nodes and the last one, are fully connected with their precedents. Let (i, j) represent a node pair, where $0 \leq i < j \leq N-1$. The core idea is to update all architecture parameters α by formulating the information propagated from i to j as a weighted sum of $|O|$ operations, where $|\cdot|$ denotes the number of operations in the set. This weighted sum f_{ij} is formulated by

$$f_{ij}(x_i) = \sum_{o \in O} \bar{\alpha}_{ij}^{o} o(x_i), \tag{1}$$

where x_i denotes the output of node i, $\bar{\alpha}_{ij}^{o}$ denotes the weight of operation o between node i and j after softmax, which is

formulated as

$$\bar{\alpha_{ij}^{o}} = \frac{\exp\left(\alpha_{ij}^{o}\right)}{\sum_{o' \in O} \exp\left(\alpha_{ij}^{o'}\right)}, \tag{2}$$

where α_{ij}^{o} is the architecture parameter of operation o between node i and j. Node j outputs the sum of all input flows, where

$$x_j = \sum_{i<j} f_{ij}\left(x_i\right). \tag{3}$$

Each cell has three types of nodes, including input nodes, inner nodes, and output nodes. The first two nodes are input nodes to a cell, and the last node, also the output of the cell, is the concatenation of inner nodes. This design allows both network parameters and architecture parameters α differentiable, so searching for the network architecture can be end-to-end.

Normal cells and reduction cells have the same micro architectures. Besides, they both use a weight-sharing strategy to reduce search overhead and help the supernet to converge. However, the architecture parameters α_n and α_r are learned separately to ensure the searched architectures of the two cells are different.

3) Micro Architecture of Learnable Stems: As illustrated in Fig. 1, both Fusing Stem and FL Stem have input nodes, inner nodes, and an output node, which are similar to the micro architectures of cells. The difference is that Fusing Stem integrates both WL and NIR fluorescence data, while FL Stem integrates NIR fluorescence data only.

Specifically, the Fusing Stem contains N_s+4 nodes, including 3 input nodes, N_s inner nodes, and 1 output node. The input nodes are C_{sin} features of WL, NIR-I, and NIR-II data after mapped by manually designed stems, where a manually designed stem has a 3×3 convolution and a batch normalization layer (BN). The inner nodes are the same as mentioned in Section II.A.2, except the last one that concatenates all precedents. The output node compresses the number of features to C_{sout} using a 1×1 convolution. The FL Stem shares the same micro architecture as Fusing Stem, except that it only has 2 input nodes.

B. Search for the Network Architecture

To search for the appropriate architectures for the multi-modal dataset, optimization strategy, loss function, and strategy to derive architectures from supernet should be considered. In DLS-DARTS, one-step gradient-based bi-level optimization [36] is used to speed up the process of updating architecture parameters. Focal loss [45] is used to alleviate the impact of imbalanced data distribution in glioma data. In addition, a modified perturbation-based derivation strategy is applied to derive the architecture with a greedy decision from supernet.

1) One-Step Gradient-Based Bi-Level Optimization: The bi-level optimization strategy is a standard searching strategy for many gradient-based NAS methods [37], [41], [46]. Our DLS-DARTS also follows this framework. In particular, the search process can be represented as a nested optimization of

$$\begin{aligned} &\min_{\alpha} L_{val}\left(w^*, \alpha\right) \\ &s.t.\ w^* = arg\min_{w} L_{train}\left(w, \alpha\right), \end{aligned} \tag{4}$$

where L_{val} denotes the loss on validation dataset, w denotes network parameters, and α denotes architecture parameters. However, the inner loop to calculate the optimal w^* consumes time, which makes the optimization slow and inefficient. To speed up the process, one-step gradient optimization is used to approximate w^*, which can be represented as

$$w^* \approx w - \xi \frac{\partial L_{train}\left(w, \alpha\right)}{\partial w}, \tag{5}$$

where ξ denotes the learning rate of w. Thus, the optimization becomes a joint optimization of α and w when using gradient descending.

2) Search for the Architecture Parameters: To achieve one-step gradient-based bi-level optimization, the original training dataset should be split into two halves. One is called training dataset for training the network parameters w, and the other is called validation dataset for updating architecture parameters α. w and α are updated alternatively each iteration.

Since the distribution of grades of intraoperative glioma data is imbalanced, focal loss [45] is introduced to reduce the impact of category imbalance. The formula of focal loss for a single sample is given by the following equation

$$FL\left(p\right) = -\sum_{i=1}^{c} \alpha_{focal}\left(1-p_i\right)^{\gamma} \cdot \left(y_i \cdot log\left(p_i\right)\right), \tag{6}$$

where FL is short for focal loss, p denotes the prediction vector of possibilities that the sample belongs to every class after softmax, p_i denotes the ith element of p, y_i denotes the ith element of the one-hot label vector, c is the number of classes, α_{focal} and γ are hyper-parameters to control the impact of label imbalance. With larger α_{focal} and γ, the impact of the minor classes to the network training will be more obvious.

To further improve the performance of DLS-DARTS, DropPath [47] during searching is also included in this study. Guo *et al.* [48] points out that the weights in the supernet are deeply coupled in weight-sharing approaches, which could mislead the architecture search process. To alleviate the weight coupling problem, DropPath randomly drops part of the operations in the cells to reduce the frequency of joint optimization of different subnetworks in the supernet.

3) Derive the Architecture From Supernet: DARTS derives the architecture from the supernet with the largest architecture parameter on each edge and keeps the two precedents with the largest $\bar{\alpha_{ij}^{o}}$ of each node. This strategy is intuitive, as architecture parameters are expected to be the weights to show the importance of the operations. However, DARTS tends to assign larger architecture parameters to skip connection, which results in worse generalization ability of derived architecture [49], [50]. Wang *et al.* [51] proposed a perturbation-based architecture selection strategy (PT) to derive the architectures that have the largest effect on the performance of the supernet. However, PT introduces random factors to the derived architectures that

might influence the performance, and the total cost of the fine-tuning processes of each trial can be expensive.

Algorithm 1 Modified Perturbation-Based Derivation Strategy (MPDS)

Input:
A pretrained supernet S
Pretrained architecture parameters $\alpha = \{\alpha_{fusing}, \alpha_{fl}, \alpha_n, \alpha_r\}$
Set of edges $\mathcal{E} = \{\mathcal{E}_{fusing}, \mathcal{E}_{fl}, \mathcal{E}_n, \mathcal{E}_r\}$ from S
precedents to keep of each node $k = \{k_{fusing}, k_{fl}, k_n, k_r\}$
Set of inner nodes $\mathcal{N} = \{\mathcal{N}_{fusing}, \mathcal{N}_{fl}, \mathcal{N}_n, \mathcal{N}_r\}$ from S
Evaluation metric m
Functions:
Onehot function $OneHot(\cdot)$
Index function $Idx(\cdot)$
Top-k worst function $TopK(\cdot, k)$
Zeroize function $Zero(\cdot)$
Output:
Derived architecture
$Geno = \{Geno_{fusing}, Geno_{fl}, Geno_n, Geno_r\}$
forall x *in* $\{fusing, fl, n, r\}$ **do**
  # Derive operation on each edge
  forall *edge* e *in* $\mathcal{E}_x$ **do**
    forall *operation on edge* o_e *in* O_e **do**
      evaluate the metric $\mathrm{m}_{\setminus o_e}$ of S when removing o_e
    end
    select o_e with the worst $\mathrm{m}_{\setminus o_e}$: $o_e^* \leftarrow arg\min_{o_e} \mathrm{m}_{\setminus o_e}$
    discretize edge e to o_e^*: $\alpha_{xe}^* \leftarrow OneHot\,(Idx\,(o_e))$
  end
  # Derive precedents of each inner node
  forall *node* n *in* $\mathcal{N}_x$ **do**
    forall *precedent* p_n *of* n **do**
      evaluate the metric $\mathrm{m}_{\setminus p_n}$ of S when removing p_n
    end
    select p_ns with the worst Top-k_x $\mathrm{m}_{\setminus p_n}$:
    $\{p_n^*\} \leftarrow TopK\left(\{\mathrm{m}_{\setminus p_n}\}, k_x\right)$
    prune out all other precedents: $\alpha_{xn}^* \leftarrow Zero\left(\{p_n\}_{\setminus\{p_n^*\}}\right)$
  end
  derive $Geno_x$ based on α_x^*
end

Our DLS-DARTS adopts PT in a sequential derivation manner, which means the derivation of operation on each edge is started from the first node to the last precedent of each node, and the operation selection is started from the first inner node to the last one. Meanwhile, the fine-tuning process is removed. Our modified perturbation-based deriving strategy (MPDS) follows the core idea of PT, which is to derive the architecture that has the most impact on the model performance in a greedy manner. Besides, removing fine-tuning on the small-scale intraoperative glioma data avoids heavy overfitting on the validation dataset, which assists the derivation process.

The modified perturbation-based algorithm works as follows: For both heads and cells, the operation selected on each edge depends on its impact on the supernet, without which the supernet will have the worst evaluation metric (e.g., accuracy). The operations are selected in a greedy manner, which means that the selection is done edge by edge. After operation selection on each edge, precedents selection of each node goes on. The number of selected precedents of each node is the same as the number of input nodes of heads or cells. The selected precedents depend on their impact on the supernet, without which the supernet will have the worst evaluation metric. Only inner nodes select precedents, except the last node of Learnable Stems. See the pseudocode of our modified perturbation-based derivation algorithm in Algorithm 1.

C. Generate the NAS-Based Model From Searched Architectures

After derivation, only the architectures of heads and cells are maintained, and a new network is constructed based on these found architectures. The constructed network is called the NAS-based model in the following. The training pipeline of the NAS-based model is the same as that of manually designed CNNs.

The NAS-based model has the same macro architecture design as DLS-DARTS, but the depth is controlled by a new parameter K_{train} rather than K_{search}, which determines the number of normal cells in each group. The micro architectures are pruned versions of the original stems and cells in DLS-DARTS.

It should be noted that the NAS-based model only inherits the architectures of stems and cells found in the training phase. The network weights are reinitialized, and architecture parameters are removed. Besides, the test dataset is appeared neither in the search phase nor the training phase. As a result, the training phase done on the original training dataset, which is the union set of training dataset and validation dataset in the search phase, does not arouse the problem of severer overfitting, and the evaluation is fair.

III. Experimental Settings

A. Data

Imaging modalities in this study included WL and fluorescence in the NIR-I and the NIR-II window. A system composed of an imaging unit and a controlling unit [21] was used to acquire all three modes of images. The imaging unit has a laser generator sub-system and a multi-spectral (visible, NIR-I, and NIR-II) imaging instrument. The laser generator sub-system generated excitation light for fluorescence imaging, with an output wavelength of 808 nm, while the visible and NIR-I/II multi-spectral imaging instrument was set to capture white light and excitation light and form images. The controlling unit was used to precisely control the working distance and reduce the touching of the imaging unit.

Indocyanine green (ICG), which is a safe NIR dye and approved by the Food and Drug Administration (FDA) for routine clinical use, was used as the imaging agent [52] for NIR imaging. Specifically, ICG was injected into patients at a dose of 1 mg/kg 48 hours before anesthesia started. Then, the multi-spectral imaging instrument was used to take multi-modal images of resected surgical specimens simultaneously during surgery. The study was approved by the Ethics Committee of Beijing Tiantan Hospital, Capital Medical University. All patients were given informed consent for their agreement. This study on patients with glioma was also explored in a clinical trial (ChiCTR2000029402) in China.

In this study, WL, NIR-I, and NIR-II images were collected from 1,115 specimens from surgery of 24 glioma patients. All specimens had all three modes of images to construct the

multi-modal dataset. Histopathologic results were used as the gold standard. Note that the specimens from one patient were not of the same grade, therefore the labels were given as the sample-level, not patient-level. Of all specimens, 302 (27.1%) were Grade I, 482 (43.2%) were Grade II, 228 (20.4%) were Grade III, and 103 (9.2%) were Grade IV. Data of 5 patients were randomly split from all to construct the test dataset (Grade I:II:III:IV=34:62:57:10), and the rest constructed the training dataset (Grade I:II:III:IV=268:420:171:93). No separate validation dataset was split from the all, and the validation dataset in the search phase was a split with half of the original training data.

The value of each pixel of fluorescence images produced by the fluorescence imaging system represented the intensity of the fluorescence signal and did not range from 0 to 255 like natural images (RGB and grayscale images). Therefore, pixel values of fluorescence images required a non-linear normalization to cope with the image classification pipeline. Besides, to reduce the influence of the inconsistent distribution of pixel values of different images caused by noise and ambient light, auto-contrast algorithms were applied to normalize the whole contrast of the images. The preprocessing was done using ImageJ [53]. In addition, to reduce the impact of noises in the fluorescence images, low-pass filtering in the frequency domain with filter size half the width and height of input resolution was applied before training and evaluation. Input data were 0-1 normalized before sending to the network.

B. Implementation Details

The operation set had 10 operations, which added 3×3 and 5×5 convolutions to the candidate operation set of DARTS. All operations were of stride one, except the ones inside reduction cells to achieve downsampling. All convolutional operations followed the order "ReLU-convolution-BN", and each separable convolution was applied twice [54], [55]. The input of each mode had 3 channels (for fluorescence images duplicated three times). The output of manually designed stems had $C_{sin} = 48$ channels during searching while $C_{sin} = 144$ during training, and the output of the learnable stems had $C_{sout} = 2 \times C_{sin}$ channels. Besides, each head had $N_s = 2$, and each cell had $N = 6$, including 2 input nodes, 3 inner fully connected nodes, and 1 channel-concatenation output node. The first two nodes were the output of the previous two cells respectively, with 1×1 convolution inserted if necessary. The number of normal cells per group K_{search} and K_{train} were both set to 1.

The input resolution of DLS-DARTS was 64×64 since images in the multi-modal intraoperative glioma dataset had similar sizes. While for comparison methods, the resolution varied according to the model structure. Besides, to increase the diversity of input data, basic data augmentation with random cropping zero-padded by 8 pixels on each edge, random horizontal flipping, and random vertical flipping, were applied both in the search and training phase. CutMix [56] with $\beta = 1.0$ was also added during the training phase. A fixed DropPath with probability 0.2 was applied during searching, while a linearly increased DropPath [47], [55] with maximum probability 0.2 was introduced during training as a regularization technique to help the network learn and prevent it from overfitting.

To search for the appropriate architecture of both heads and cells, epochs of 50, batch size of 16, learning rate for network parameters of 2.5×10^{-3}, and architecture learning rate for architecture α of 6×10^{-4} were initially set. SGD optimizer with momentum 0.9, weight decay of 3×10^{-4}, and a cosine annealing schedule down to 10^{-5} was used for updating network parameters, and Adam optimizer with weight decay of 1×10^{-3}, β_1 of 0.5, and β_2 of 0.999 was used for updating architecture α. After obtaining the architecture of cells, we changed epoch to 110 and batch size to 128 in the training phase for better training. SGD optimizer with momentum 0.9, weight decay of 1×10^{-5}, and a cosine annealing schedule down to 10^{-5} was used for training the NAS-based model. Focal loss with $\gamma = 2.0$ and $\alpha = 1.0$ was used both in searching and training. Hyperparameters of DLS-DARTS and other methods for comparison were tuned to the best. All the evaluations were conducted on the test dataset.

In this work, Google Colab TPU Runtime (v2, 8 Cores, 64 GB Memory) was used under the environment of Python 3.6 to perform the experiments of all grading models. TPU was used for speeding up the training process, including the search phase of NAS models and training of all models. TensorFlow [57] (version 2.3) was used for image preprocessing, network building, training, and evaluation.

C. Evaluation Metrics

To evaluate the performance of all the classification models, accuracy (ACC) and AUC were used as the evaluation metrics. AUC is the area under the receiver operating characteristic (ROC) curve, which takes true positive rate (TPR) as the y-axis and false positive rate (FPR) as the x-axis. The metrics are defined as:

$$ACC = \frac{\sum_{i=1}^{c} TP_i}{\#TotalSamples}, \tag{7}$$

$$TPR = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} (TP_i + FN_i)}, \tag{8}$$

$$FPR = \frac{\sum_{i=1}^{c} FP_i}{\sum_{i=1}^{c} (FP_i + TN_i)}, \tag{9}$$

$$AUC = \int_0^1 f(x)\, dx, \tag{10}$$

where c is number of grades, $\#Total\ Samples$ is the number of samples for evaluation, TP_i is the true positive number of the ith class, TN_i is the true negative number of the ith class, FP_i is the false positive number of the ith class, FN_i is the false negative number of the ith class, and $f(x)$ is the corresponding TPR when FPR is x. In this study, TP, TN, FP, FN, and AUC were calculated using a micro average that treated each element of the label indicator matrix as a label [58].

IV. Results

A. Comparison of NAS With Manually Designed CNNs

To evaluate the effectiveness of DLS-DARTS, here four methods in classification tasks were introduced for

comparison, including three manually designed CNNs ResNet [59], PyramidNet [60] with ShakeDrop regularization [61], and EfficientNet [62], and a state-of-the-art NAS-based method MMNAS [41] for PET-CT multi-modal classification in medical image analysis.

ResNet is one of the most classical CNN models, and the proposed residual connection has a deep influence on the design of the model architecture afterward. It has been used as a baseline in MedMNIST [63], a classification benchmark based on medical images for testing NAS methods. ResNet-18 with input resolution 224 has shown competitive results to NAS methods on BreastMNIST [64], a dataset that also has features of low-resolution, heavy noise, and small scale. EfficientNet strikes a balance between speed and performance through adjusting the depth, width, and resolution of the model carefully, and becomes one of the state-of-the-art CNNs in the classification task. It has also been used to analyze the boundary of glioma intraoperatively [22]. Besides, the weights pretrained on ImageNet [65], a large-scale natural image dataset, are publicly available for transfer learning.[1] PyramidNet with ShakeDrop regularization is one of the state-of-the-art methods on CIFAR [66], a natural image dataset that also has the feature of low-resolution. Since these manually designed CNNs have shown remarkable results on tasks that have similar features to the intraoperative glioma dataset, we selected them as comparisons to evaluate the effectiveness of our DLS-DARTS.

MMNAS is a state-of-the-art DARTS-based method for the classification of soft-tissue sarcomas on PET-CT images. The method shows many similarities to our DLS-DARTS, with a normal cell at the beginning to merge the low-level features of PET and CT and the gradient-based updating strategy to search for the architectures, thus also being compared with our DLS-DARTS. To adapt this method in our multi-modal intraoperative glioma images, we modified the elementwise sum of two input modes to concatenation, because PET and CT are complementary but optical imaging modalities are not. We also added a stem for fluorescence images. The NIR-I and NIR-II images were first concatenated and then be seen as an entirety in the network.

DLS-DARTS was compared with ResNet-18, EfficientNet-B0, EfficientNet-B0 with transfer learning, PyramidNet-ShakeDrop, and MMNAS to show its advantages. Note that the architecture of the MMNAS-based model was searched using our multi-modal dataset. All manually designed CNNs had an input channel of 9, which concatenated the images of WL, NIR-I, and NIR-II. Although EfficientNet-B0 had the least number of parameters and the simplest architecture in the EfficientNet family, it outperformed B1 to B3 in intraoperative glioma grading. This result might be caused by the small resolution of the dataset that did not match the design of larger models in the family. Since larger EfficientNets required larger input resolution, the input images required to be scaled to a larger size by interpolation. Thus, useful information contained in the original images became sparse in larger resolutions, which possibly made the models focus more on the interpolated redundant information. The low-level features of original images were desalted and the performance of larger models was hurt. For transfer learning based on EfficientNet-B0, the stem of the transferred model was dropped out and reinitialized to match the input with 9 channels. All the input images applied data augmentations dynamically, so every epoch the input from the same original image was different. Due to the CNNs being fine-tuned to their best state, the comparison was relatively fair. The search phase of DLS-DARTS cost 0.75 GPU hours on a single Tesla P100. The searched architectures of both learnable stems and cells are shown in Fig. 2. Note that the derived architecture had 3×3 convolution operations inside (nor_conv_3 $\times$ 3 in the figure), indicating that the expansion of the search space for candidate operations was effective.

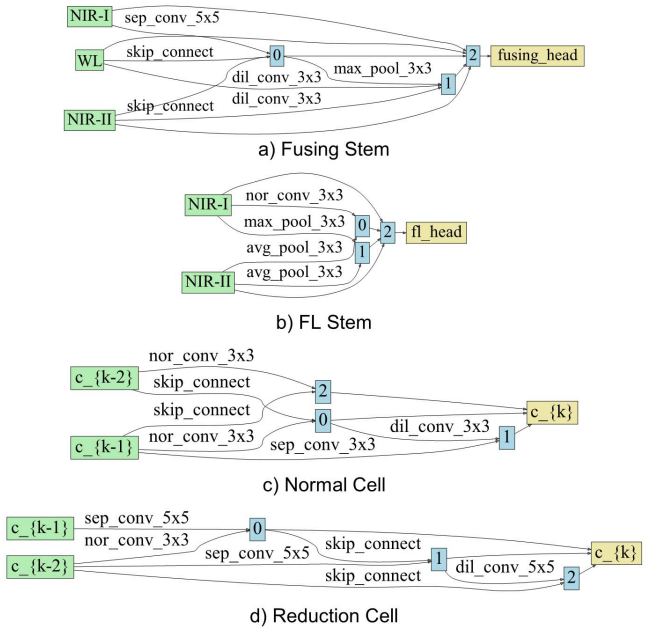
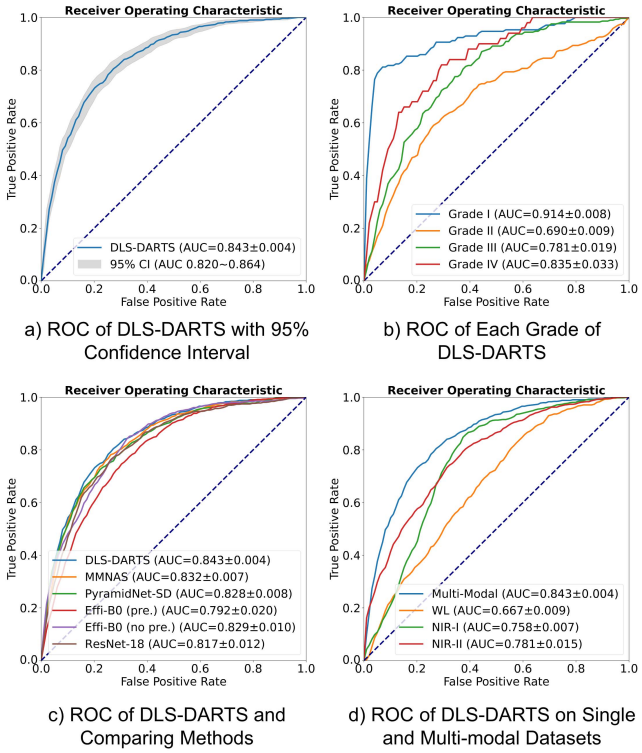


Fig. 2. The architecture of the discovered Fusing Stem, FL Stem, normal cells, and reduction cells of DLS-DARTS on multi-modal intraoperative glioma dataset. Nodes represent features in the network. nor_conv_3 $\times$ 3 denotes 3 $\times$ 3 convolution operation.

The results of DLS-DARTS and comparing methods evaluated on the test dataset on 5 runs are shown in Table I. The ROC curve of DLS-DARTS with 95% Confidence Interval (CI) and ROC curves of each grade of DLS-DARTS is shown in Fig. 3 a) and b), respectively. The comparison of the ROC curves of DLS-DARTS and comparing methods are shown in Fig. 3 c). The number of parameters, the number of floating-point operations (FLOPs), and throughput were used to evaluate the complexity, computation of models, and inference time, respectively. DLS-DARTS achieved the best ACC (0.634, 95% CI 0.602∼0.669) and AUC (0.843, 95% CI 0.820∼0.864), which outperformed MMNAS and manually designed CNNs. Besides, DLS-DARTS showed remarkable ability to distinguish Grade I, but was weaker for higher grades. This might be due to the diffusion characteristics of glioma of Grade II to IV that increase the difficulty in distinguishing the details of the tissues. In addition, the

[1] https://github.com/tensorflow/tensorflow



a) ROC of DLS-DARTS with 95% Confidence Interval

b) ROC of Each Grade of DLS-DARTS

c) ROC of DLS-DARTS and Comparing Methods

d) ROC of DLS-DARTS on Single and Multi-modal Datasets

Fig. 3. The ROC curves of a) DLS-DARTS with 95% confidence interval (CI) of DLS-DARTS; b) each grade of DLS-DARTS; c) DLS-DARTS and comparing methods; d) DLS-DARTS on single-modal and multi-modal datasets. The confidence interval was calculated based on 1000 iterations of the bootstrap methods. All the ROC curves were drawn on the concatenation of the results of 5 runs using the micro average. All AUCs (mean±std) were evaluated on 5 runs.

TABLE I

PERFORMANCE OF DLS-DARTS AND COMPARING METHODS INCLUDING MMNAS AND MANUALLY DESIGNED CNNS AVERAGED FROM 5 RUNS. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD. SD IS SHORT FOR SHAKEDROP. THROUGHPUT WAS EVALUATED ON A SINGLE TESLA P100 GPU AVERAGED FROM 50 BATCHES WITH BATCH SIZE OF 128

Model	# Params (M)	# FLOPs (G)	Through-put (im/s)	ACC	AUC
DLS-DARTS	5.11	1.99	504	**0.634** (0.602,0.669)	**0.843** (0.820,0.864)
MMNAS	41.17	7.41	336	0.597 (0.564,0.632)	0.832 (0.808,0.854)
PyramidNet-SD	27.21	18.49	96	0.608 (0.574,0.644)	0.828 (0.805,0.851)
EfficientNet-B0 (pretrained)	4.06	0.38	1315	0.527 (0.494,0.563)	0.792 (0.768,0.816)
EfficientNet-B0 (not pretrained)	4.06	0.38	1315	0.596 (0.563,0.629)	0.829 (0.801,0.846)
ResNet-18	11.21	2.06	2172	0.602 (0.569,0.637)	0.817 (0.792,0.840)

number of parameters of DLS-DARTS was only more than EfficientNet-B0, indicating the effectiveness of the parameters in the model.

TABLE II

PERFORMANCE OF DLS-DARTS USING DIFFERENT MODES OF IMAGES AVERAGED FROM 5 RUNS. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRAILS OF THE BOOTSTRAP METHOD

Image Mode	# Params (M)	# FLOPs (G)	ACC	AUC
Multi-modal	5.11	1.99	**0.634** (0.602,0.669)	**0.843** (0.820,0.864)
WL	6.21	3.01	0.369 (0.335,0.402)	0.667 (0.636,0.692)
NIR-I	7.35	2.64	0.425 (0.388,0.456)	0.758 (0.732,0.781)
NIR-II	2.84	2.28	0.549 (0.517,0.584)	0.781 (0.756,0.805)

Although MMNAS was also a gradient-based NAS method for multi-modal medical image analysis, it was designed for PET-CT rather than intraoperative optical images. Besides, the normal cell in MMNAS also fused low-level features of different modes, which was similar to our learnable stems. However, the remaining structures of MMNAS were only reduction cells, which limited the diversity of the network, and further limited its ability to fuse and extract features in deeper layers. These results explained why our DLS-DARTS outperformed MMNAS. PyramidNet-ShakeDrop showed competitive performance on AUC to DLS-DARTS, but the network structure was obviously more complex, and the number of FLOPs was ∼10 times of DLS-DARTS. EfficientNet-B0 had a balanced performance on network characteristics and evaluation metrics, but the ACC is worse than DLS-DARTS. Although DLS-DARTS had obviously larger FLOPs and the throughput was only half of EfficientNet-B0, it met the demand for real-time diagnosis. ResNet-18 had similar FLOPs to DLS-DARTS, but with an obvious AUC gap. Interestingly, pretrained EfficientNet-B0 showed worse performance compared to the one without pretrained weights. This might be explained by that the multi-modal optical images had obvious differences over natural images, therefore transfer learning strategy showed poor improvement in intraoperative glioma grading. Similar results are reported in [67]. Besides, since the weight of the pretrained stem was dropped out to match the multi-modal dataset, the ability of the model to extract low-level features was severely influenced. This change also caused the performance gap. The results also indicated the importance of low-level features in distinguishing the grades of intraoperative optical glioma images. Meanwhile, the ROC curve of DLS-DARTS was the nearest to the upper left corner, which indicated the best overall prediction performance as well.

B. Comparison of Single-Modal and Multi-Modal Imaging

To study the potential advantages of multi-modal imaging, we then evaluated DLS-DARTS on both single-modal images and multi-model images. Models of each mode shared the same hyperparameters for a fair comparison.

The results of the four models are shown in Table II, and the comparison of ROC curves is shown in Fig. 3 d).

The model trained on multi-modal images achieved the best ACC (0.634, 95% CI 0.602∼0.669) and AUC (0.843, 95% CI 0.820∼0.864), which had a significant advantage over the other three models trained on single-modal images. Notably, the model trained on multi-modal images also had the least number of FLOPs (1.99 G). This phenomenon indicated that multi-modal images contained richer information than single-modal images, thus the model could capture sufficient useful features with fewer calculations. Interestingly, the NAS-based model searched from NIR-II images had the least number of parameters. This indicated that NIR-II imaging might provide more information than other modes, thus requiring less parametric operations to extract the features. Experimental results demonstrated that multi-modal images contained more information for the NAS- based model to learn than single-modal images. We also noted that the models searched on fluorescence images generally performed better than the model searched on WL images, while the model searched on NIR-II images performed significantly better than the ones on NIR-I and WL. This indicated that fluorescence imaging could better reflect some features of biological tissues, especially NIR-II imaging that showed advantages over NIR-I imaging to reflect these features.

C. Gradient-Weighted Class Activation Mapping

Gradient-weighted class activation mapping (GCAM) [68], known as a technology for visualizing and explaining the decision of CNN models, was also implemented in this study for a better understanding of DLS-DARTS. This technology calculated the contributions of different parts in an image when the model was finally predicting a specific class. Heatmap was used to display the attention of the searched model, with red illustrating the most concerning part, while dark blue the least.

After the training phase of DLS-DARTS, some multi-modal images were randomly picked up with different grades. GCAM was implemented to check the model's focusing part when predicting the grades of these images. Since the multi-modal images could not be displayed directly, we split the input images and mixed the heatmap of their specific grades. WL images were selected as representations of mixed images to display the result in a salience way in Fig. 4. Here, the fluorescence images were normalized to range 0 to 255, so they could be seen directly. The result showed that the NAS model automatically found out the important and distinguishable parts on specimens that could hardly be recognized by human eyes, which might explain its ability to determine the grades of different inputs.

D. Ablation Studies for Analyzing DLS-DARTS

To further understood how every part of DLS-DARTS influenced its performance, ablation studies on strategies for derivation and the learnable stems were conducted. All the hyperparameters were the same as the ones in Section IV.A, and the evaluation was performed on the test dataset. For the study of derivation strategy, the architectures were searched separately. However, for studies of learnable stems, we used the same architectures in Fig. 2. The input features of the ablated stems were concatenated and a 1×1 convolution was performed afterward to keep the channels the same as the original DLS-DARTS. The results are shown in Table III. Both MPDS and the two learnable stems contributed to the good performance on AUC and ACC. The results indicated the novelty of our DLS-DARTS in intraoperative glioma grading using multi-modal optical imaging.

TABLE III
ABLATION STUDIES OF DLS-DARTS ON DERIVATION STRATEGY AND LEARNABLE STEMS AVERAGED FROM 5 RUNS. THE ABLATION PARTS WERE REMOVED FROM DLS-DARTS. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRAILS OF THE BOOTSTRAP METHOD. Δ IS THE INCREMENT OF AVERAGE PERFORMANCE. MPDS DENOTES THE MODIFIED PERTURBATION-BASED DERIVATION STRATEGY

Ablation Part	ACC	ΔACC	AUC	ΔAUC
DLS-DARTS	0.634 (0.602,0.669)	0	0.843 (0.820,0.864)	0
- MPDS	0.616 (0.574,0.643)	-0.018	0.835 (0.811,0.855)	-0.008
- Fusing Stem	0.607 (0.574,0.640)	-0.027	0.832 (0.809,0.855)	-0.011
- FL Stem	0.600 (0.567,0.632)	-0.034	0.831 (0.811,0.855)	-0.012
- Learnable Stems	0.573 (0.539,0.606)	-0.061	0.829 (0.805,0.850)	-0.014

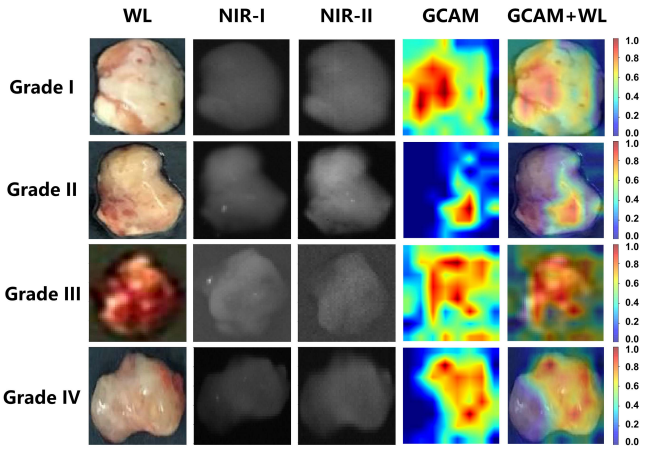


Fig. 4. The GCAM results of the NAS-based model on multi-modal intraoperative glioma images. The first three columns are WL, NIR-I and NIR-II images, respectively. The GCAM column is the heatmap that displays the attention of the model when making the final prediction of a specific grade. The last column is the mixture of GCAM and the original WL image.

E. Impact of Hyperparameters for Supernet Design

To explore the impact of hyperparameters for supernet design, we evaluated three important hyperparameters, including the selection of candidate operation set, the number of normal cells per group, and the number of total reduction cells. We evaluated the AUC and ACC to show the impact.

In particular, we evaluated the following three candidate operation sets of a) DARTS (8 operations); b) DARTS without

TABLE IV
IMPACT OF DIFFERENT CANDIDATE OPERATION SETS AVERAGED FROM 5 RUNS. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

O	$\|O\|$	ACC	AUC
DARTS	8	0.580 (0.546,0.613)	0.825 (0.786,0.834)
DARTS -{5x5 dil conv} +{1×1, 3×3 nor_conv}	9	0.600 (0.566,0.633)	0.831 (0.809,0.855)
DARTS +{3×3, 5×5 nor_conv}	10	**0.634** (0.602,0.669)	**0.843** (0.820,0.864)

TABLE V
IMPACT OF DIFFERENT NUMBERS OF NORMAL CELLS PER GROUP AVERAGED FROM 5 RUNS. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

K_{train}	L	ACC	AUC
1	6	**0.634** (0.602,0.669)	**0.843** (0.820,0.864)
2	9	0.606 (0.547,0.639)	0.832 (0.811,0.855)
3	12	0.617 (0.585,0.650)	0.820 (0.799,0.845)

TABLE VI
IMPACT OF DIFFERENT NUMBERS OF TOTAL REDUCTION CELLS AVERAGED FROM 5 RUNS. NUMBERS IN BRACKETS INDICATE 95% CONFIDENCE INTERVALS USING 1000 TRIALS OF THE BOOTSTRAP METHOD

K_r	L	ACC	AUC
2	4	0.582 (0.548,0.616)	0.826 (0.803,0.848)
3	6	**0.634** (0.602,0.669)	**0.843** (0.820,0.864)
4	8	0.590 (0.557,0.625)	0.831 (0.809,0.853)

5×5 dilated convolution but with 1×1 and 3×3 convolutions (9 operations), and; c) DARTS with 3×3 and 5×5 convolutions (10 operations), which was also the candidate operation set of MMNAS with 2D convolutions. Set b) had a receptive field no larger than 5 for a single operation, while the added convolutions had competitive parameters compared to convolutions already in the set. As shown in Table IV, set c) showed the best performance and was chosen in our experiments. The superiority of set c) might be contributed to the normal convolution and operations with a larger receptive field.

We then set the number of normal cells K_{train} to 1, 2, and 3, respectively. The results are shown in Table V, where L denotes the total number of cells in DLS-DARTS. The searched model yielded the best performance when $K_{train} = 1$. The experimental results showed that a shallow model was enough for the multi-modal intraoperative glioma grading, which might be due to that the model was easier to get overfitted on the small-scale dataset as the depth of the model increased.

Finally, we evaluated the impact of the number of total reduction cells K_r. We set K_r to 2, 3, and 4, respectively. The input resolution of all models was set to 64×64. The results are shown in Table VI. The NAS-based model performed the best when $K_r = 3$, with feature size 8×8 before global average pooling. The results are consistent with classification models designed for small-resolution natural images like CIFAR, indicating the proper size of features that contain beneficial information to distinguish different grades.

V. DISCUSSION AND CONCLUSION

In this study, we develop a NAS method DLS-DARTS for intraoperative glioma grading using multi-modal imaging. The multi-modal dataset is constructed by WL, NIR-I, and NIR-II images that are simultaneously obtained during surgery. Our method achieves the best performance, proving the effectiveness and advancement of automatically searched network architectures over manually designed networks for intraoperative glioma grading. It also gains significant improvement over models trained on single-modal imaging, showing its high potential to be used in real-time diagnosis.

There are three major frameworks in NAS methods, including evolution algorithms, reinforcement learning, and gradient optimization. Compared with evolution algorithms and reinforcement learning, gradient optimization is more efficient and simpler to be used. Although evolution algorithms and reinforcement learning show higher upper bounds for the performance in natural image classification [68], they usually require a huge amount of GPU time compared to gradient optimization, which limits their wide applications in new areas such as medical image analysis. DARTS, as the most classical gradient optimization method, has been used in medical image analysis and shows compelling performance [41], [70]. Therefore, we adopted DARTS as our main architecture to achieve real-time intraoperative glioma grading.

Since NAS models are explored from a large and complicated search space that may be hard for humans to imagine, there exist chances for good network architectures that are superior to currently designed manual CNNs. Besides, NAS provides a method to explore architectures rather than offering concrete architectures, thus the NAS-explored architectures may be better-suited to the target task compared with manually designed CNNs. These advantages of NAS motivate us to use this powerful technique to deal with challenges caused by the deficiencies of NIR images, and fully utilize multi-modal information. Our work demonstrates that DLS-DARTS can extract features from multi-modal images more effectively, which yields better performance on AUC and ACC compared with manually designed CNNs. In addition, the comparison between DLS-DARTS trained on multi-modal images and single-modal images proves that WL, NIR-I, and NIR-II images contain different and complementary features for models to learn. Multi-modal imaging indeed provides more abundant information than single-modal imaging and boosts the performance of DLS-DARTS. These results show great prospects of NAS in medical image analysis. We believe

NAS can be used as a new pipeline to cope with difficulties in more situations such as lesion segmentation and pathological analysis.

Previous works on glioma grading mainly focus on patient-level preoperative imaging and postoperative pathology. They build automated pipelines for diagnosis and treatment planning, aiming at saving time, resources, and labor. Unlike these works, we concentrate on the real-time sample-level diagnosis of glioma during surgery for precise resection guidance. DLS-DARTS can grade one specimen into four grades within 1 s, which guarantees both low latency and high throughput. Therefore, it can provide diagnoses for even tens or hundreds of specimens that better meet the demand of neurosurgeons during surgery. It also has the potential to help decide the dosage of drugs and radiation early after surgery. Compared with neurosurgeons who fail to grade the glioma tissues using WL/NIR-I/NIR-II images, our DLS-DARTS provides a new pipeline to grade the glioma tissues during surgery. While compared with conventional intraoperative pathology, our method is much simpler and does not require pathologist for manual grading. It also avoids sampling error of intraoperative frozen sections [71] since the diagnosis is in real-time. However, we should admit that the accuracy currently DLS-DARTS achieves is a bit lower than commonly used frozen sections in the clinical practice [72], [73]. It warrants for further improvement.

This work shows the potential of NAS for intraoperative glioma grading. However, there are some limitations as well. NAS does not provide concrete architectures before searching, which increases the difficulty of understanding it. The architectures searched by NAS are also complicated compared with manually designed CNNs. As a result, people tend to use manually designed CNNs that are easier to tune rather than NAS when they deal with new tasks. Besides, DLS-DARTS is sensitive to hyperparameters for both search and training on small-scale datasets, which may limit its generalization performance. The intraoperative imaging techniques also need further improvement for better performance. In addition, multi-modal imaging is more difficult than single-modal imaging due to higher acquisition costs and requirements for advanced equipment.

To overcome these drawbacks, we plan to collect a larger dataset from multiple centers to improve the generalization performance of the model, and better assess the value of our method in clinical practice. Using a black box or keeping imaging equipment in a low-temperature environment when imaging might reduce negative environmental effects and improve imaging quality. Besides, since the area of NAS evolves fast, advanced NAS methods might be tried to search for architectures that can better extract features from the datasets.

Acknowledgment

The authors would like to thank the instrumental and technical support of the multi-modal biomedical imaging experimental platform, Institute of Automation, Chinese Academy of Sciences and also would like to thank Shuo Wang for his help.

References

[1] S. Lapointe, A. Perry, and N. A. Butowski, "Primary brain tumours in adults," *Lancet*, vol. 392, no. 10145, pp. 432–446, Aug. 2018.

[2] Q. T. Ostrom *et al.*, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016," *Neuro-Oncol.*, vol. 21, no. 5, pp. v1–v100, 2019.

[3] 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.

[4] W. Wu *et al.*, "Joint NCCTG and NABTC prognostic factors analysis for high-grade recurrent glioma," *Neuro-Oncol.*, vol. 12, no. 2, pp. 164–172, Feb. 2010.

[5] A. M. Molinaro *et al.*, "Association of maximal extent of resection of contrast-enhanced and non-contrast-enhanced tumor with survival within molecular subgroups of patients with newly diagnosed glioblastoma," *JAMA Oncol.*, vol. 6, no. 4, pp. 495–503, 2020.

[6] T. C. Hollon *et al.*, "Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks," *Nature Med.*, vol. 26, no. 1, pp. 52–58, Jan. 2020.

[7] H. Guo, J. Yu, Z. Hu, H. Yi, Y. Hou, and X. He, "A hybrid clustering algorithm for multiple-source resolving in bioluminescence tomography," *J. Biophotonics*, vol. 11, no. 4, Apr. 2018, Art. no. e201700056.

[8] S. Hu, H. Kang, Y. Baek, G. El Fakhri, A. Kuang, and H. S. Choi, "Real-time imaging of brain tumor for image-guided surgery," *Adv. Healthcare Mater.*, vol. 7, no. 16, Aug. 2018, Art. no. 1800066.

[9] D. Maric *et al.*, "Whole-brain tissue mapping toolkit using large-scale highly multiplexed immunofluorescence imaging and deep neural networks," *Nature Commun.*, vol. 12, no. 1, p. 1550, Dec. 2021.

[10] C. Qin, J. Zhong, Z. Hu, X. Yang, and J. Tian, "Recent advances in cerenkov luminescence and tomography imaging," *IEEE J. Sel. Topics Quantum Electron.*, vol. 18, no. 3, pp. 1084–1093, May 2012.

[11] Z. Zhang, M. Cai, C. Bao, Z. Hu, and J. Tian, "Endoscopic cerenkov luminescence imaging and image-guided tumor resection on hepatocellular carcinoma-bearing mouse models," *Nanomed., Nanotechnol., Biol. Med.*, vol. 17, pp. 62–70, Apr. 2019.

[12] T. Song *et al.*, "A novel endoscopic Cerenkov luminescence imaging system for intraoperative surgical navigation," *Mol. Imag.*, vol. 14, no. 8, pp. 443–451, 2015.

[13] H. Liu *et al.*, "Multispectral hybrid Cerenkov luminescence tomography based on the finite element SPN method," *J. Biomed. Opt.*, vol. 20, no. 8, Aug. 2015, Art. no. 086007.

[14] Z. Hu *et al.*, "*In vivo* nanoparticle-mediated radiopharmaceutical-excited fluorescence molecular imaging," *Nature Commun.*, vol. 6, no. 1, p. 7560, 2015.

[15] Z. Hu *et al.*, "Experimental Cerenkov luminescence tomography of the mouse model with SPECT imaging validation," *Opt. Exp.*, vol. 18, no. 24, pp. 24441–24450, 2015.

[16] B. Chang *et al.*, "A phosphorescent probe for *in vivo* imaging in the second near-infrared window," *Nature Biomed. Eng.*, 2021, doi: 10.1038/s41551-021-00773-2.

[17] M. Liu *et al.*, "Cerenkov luminescence imaging on evaluation of early response to chemotherapy of drug-resistant gastric cancer," *Nanomed., Nanotechnol., Biol. Med.*, vol. 14, no. 1, pp. 205–213, Jan. 2018.

[18] Y. Suo *et al.*, "NIR-II fluorescence endoscopy for targeted imaging of colorectal cancer," *Adv. Healthcare Mater.*, vol. 8, no. 23, Dec. 2019, Art. no. 1900974.

[19] Z. Hu, W.-H. Chen, J. Tian, and Z. Cheng, "NIRF nanoprobes for cancer molecular imaging: Approaching clinic," *Trends Mol. Med.*, vol. 26, no. 5, pp. 469–482, May 2020.

[20] Q.-Y. Chen *et al.*, "Safety and efficacy of indocyanine green tracer-guided lymph node dissection during laparoscopic radical gastrectomy in patients with gastric cancer: A randomized clinical trial," *Jama Surg.*, vol. 155, no. 4, pp. 300–311, 2020.

[21] Z. Hu *et al.*, "First-in-human liver-tumour surgery guided by multispectral fluorescence imaging in the visible and near-infrared-I/II windows," *Nature Biomed. Eng.*, vol. 4, no. 3, pp. 259–271, Mar. 2020.

[22] B. Shen *et al.*, "Real-time intraoperative glioma diagnosis using fluorescence imaging and deep convolutional neural networks," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 48, no. 11, pp. 3482–3492, Oct. 2021.

[23] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Med. Phys.*, vol. 44, no. 2, pp. 547–557, 2017.

[24] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, Apr. 2020.

[25] T.-C. Chiang, Y.-S. Huang, R.-T. Chen, C.-S. Huang, and R.-F. Chang, "Tumor detection in automated breast ultrasound using 3-D CNN and prioritized candidate aggregation," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 240–249, Jan. 2019.

[26] A. Hekler *et al.*, "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *Eur. J. Cancer*, vol. 118, pp. 91–96, Sep. 2019.

[27] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, 2016.

[28] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, 2019.

[29] N. Coudray *et al.*, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[30] C. Lin *et al.*, "CIR-Net: Automatic classification of human chromosome based on inception-resnet architecture," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jun. 18, 2020, doi: 10.1109/TCBB.2020.3003445.

[31] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, "Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma," *Sci. Rep.*, vol. 7, no. 1, p. 5467, Dec. 2017.

[32] J. Ker, Y. Bai, H. Y. Lee, J. Rao, and L. Wang, "Automated brain histology classification using machine learning," *J. Clin. Neurosci.*, vol. 66, pp. 239–245, Aug. 2019.

[33] A. Yonekura, H. Kawanaka, V. B. S. Prasath, B. J. Aronow, and H. Takase, "Automatic disease stage classification of glioblastoma multiforme histopathological images using deep convolutional neural network," *Biomed. Eng. Lett.*, vol. 8, no. 3, pp. 321–327, Aug. 2018.

[34] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.

[35] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. discovery data mining*, Aug. 2013, pp. 847–855.

[36] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," 2018, *arXiv:1806.09055*.

[37] Y. Xu *et al.*, "PC-DARTS: Partial channel connections for memory-efficient architecture search," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–13.

[38] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[39] L.-C. Chen *et al.*, "Searching for efficient multi-scale architectures for dense image prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 8713–8724.

[40] Q. Huang, Y. Xian, P. Wu, J. Yi, H. Qu, and D. Metaxas, "Enhanced MRI reconstruction network using neural architecture search," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Lima, Peru: Springer, 2020, pp. 634–643.

[41] Y. Peng, L. Bi, M. Fulham, D. Feng, and J. Kim, "Multi-modality information fusion for radiomics-based neural architecture search," in *Int. Conf. Med. Image Comput.-Assist. Intervent.* Springer, 2020, pp. 763–771.

[42] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-Unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019.

[43] Q. Zuo, J. Zhang, and Y. Yang, "DMC-fusion: Deep multi-cascade fusion with classifier-based feature synthesis for medical multi-modal images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3438–3499, May 2021.

[44] X. He, Y. Deng, L. Fang, and Q. Peng, "Multi-modal retinal image classification with modality-specific attention network," *IEEE Trans. Med. Imag.*, vol. 40, no. 6, pp. 1591–1602, Jun. 2021.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[46] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–13.

[47] G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-deep neural networks without residuals," *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–11.

[48] Z. Guo *et al.*, "Single path one-shot neural architecture search with uniform sampling," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 544–560.

[49] T. E. Arber Zela, T. Saikia, Y. Marrakchi, T. Brox, and F. Hutter, "Understanding and robustifying differentiable architecture search," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–28.

[50] L. Li and A. Talwalkar, "Random search and reproducibility for neural architecture search," in *Proc. 35th Uncertainty Artif. Intell. Conf.*, vol. 115, Jul. 2020, pp. 367–377.

[51] R. Wang, M. Cheng, X. Chen, X. Tang, and C.-J. Hsieh, "Rethinking architecture selection in differentiable NAS," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–18.

[52] J. Y. K. Lee *et al.*, "Intraoperative near-infrared optical imaging can localize gadolinium-enhancing gliomas during surgery," *Neurosurgery*, vol. 79, no. 6, pp. 856–871, Dec. 2016.

[53] M. D. Abràmoff, P. J. Magalhães, and S. J. Ram, "Image processing with ImageJ," *Biophoton. Int.*, vol. 11, no. 7, pp. 36–42, 2004.

[54] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 4780–4789.

[55] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[56] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[57] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

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

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

[60] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6307–6315.

[61] Y. Yamada, M. Iwamura, T. Akiba, and K. Kise, "Shakedrop regularization for deep residual learning," *IEEE Access*, vol. 7, pp. 186126–186136, 2019.

[62] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[63] J. Yang, R. Shi, and B. Ni, "MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 191–195.

[64] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[65] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[66] A. Krizhevsky, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.222.9220

[67] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019.

[68] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.

[69] P. Ren *et al.*, "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Surv.*, vol. 54, no. 4, p. 76, 2021.

[70] X. Yan, W. Jiang, Y. Shi, and C. Zhuo, "MS-NAS: Multi-scale neural architecture search for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Springer, 2020, pp. 388–397.

[71] R. Dammers, J. W. Schouten, I. K. Haitsma, A. J. P. E. Vincent, J. M. Kros, and C. M. F. Dirven, "Towards improving the safety and diagnostic yield of stereotactic biopsy in a single centre," *Acta Neurochirurgica*, vol. 152, no. 11, pp. 1915–1921, Nov. 2010.

[72] L. Di *et al.*, "Stimulated Raman histology for rapid intraoperative diagnosis of gliomas," *World Neurosurg.*, vol. 150, pp. e135–e143, Jun. 2021.

[73] S. Jain, M. Kaushal, A. Choudhary, and M. Bhardwaj, "Comparative evaluation of squash smear and frozen section in the intraoperative diagnosis of central nervous system tumours," *Cytopathology*, vol. 33, no. 1, pp. 107–113, Jan. 2022.