User-Assisted Design of a Neural Network for Brain Tumor Segmentation

Matheus A. Cerqueira
Institute of Computing
University of Campinas
Campinas, Brazil
Email: matheus.cerqueira@students.ic.unicamp.br

Alexandre X. Falcão
Institute of Computing
University of Campinas
Campinas, Brazil
Email: afalcão@ic.unicamp.br

***Abstract*—Brain tumor segmentation is a complicated task, with deep learning (DL) presenting the best results. However, DL segmentation models have been increasing in complexity over the last few years, requiring a high volume of fully-annotated images, which is aggravated by the fact that those models are trained to optimize a loss function with no or less understanding of how the features are learned. In contrast, a recent methodology, Feature Learning from Image Markers (FLIM), has involved an expert in the learning loop while reducing human effort in data annotation without using the backpropagation algorithm. In this work, We employ a method for estimating and selecting filters that explore user knowledge, ensuring that the first convolutional layer has features that activate the different lesions and healthy tissue patterns. We used a small U-shaped network (sU-Net) where the encoder is trained with two FLIM modifications, first with multiple training steps (MS-FLIM) and the second by using distinct configurations of markers and images using a biased FLIM (B-FLIM). The results show that the sU-Net based on MS-FLIM and B-FLIM outperforms the standard FLIM and the backpropagation algorithm. Also, We showed that our methodology achieves effectiveness within the standard deviations of the SOTA models while using a small number of layers.**

I. Introduction

Gliomas are the most common type of brain tumor in adults, with the highest occurrence in brain regions, although other areas of the Central Nervous System (CNS) might be affected [1]. Among them are Glioblastoma (GBM), the most common malignant brain tumor of the Central Nervous System (CNS) in Adults. In 2019 in the United States (US), the survival rate within five years after diagnosis was only 6.9%, with an incidence rate of 2.55 per 100,000 people [2].

The use of images is important for the initial diagnosis, with volume estimation essential for monitoring, investigating tumor progression, and analyzing the selected treatment [3], [4]. However, manual annotation is time-consuming, tedious, and error-prone – facts that have motivated research on automatic and semi-automatic methods for brain tumor segmentation.

Deep Learning (DL) presents the best results among automatic Brain Tumor Segmentation (BTS) techniques. However, DL training requires a high volume of fully-labeled images, powerful hardware to support volumetric images, and an elevated training time. Besides that, DL models for BTS have become more complex, either by increasing the number of layers or by using different models together as cascaded models, significantly increasing the number of parameters. [5]–[8].

Also, the training of those models, in general, is done by backpropagating the error from a given loss function of the desired and predicted output. In other words, there is no control or understanding of the learned features from the model. For example, in a classification task, there is no understanding if the model predicts a fish from an image because it is really a fish image or if the image has an underwater context [9].

In that sense, there is a need for greater responsibility and security in automatic or computer-aided systems in healthcare, which is reflected in work towards more transparent models with a certain level of explainability/interpretability. A solution for this problem is to use expert knowledge to build more safe and explainable models, which can also increase their performance [10].

Developed by Italos et al., a recent technique called Feature Learning from Image Markers (FLIM) reduces the human effort to mark representative class regions in fewer images. With each marked region as a candidate filter, FLIM learns convolutional filters directly from those marked regions [11], [12].

FLIM differs from traditional scribble learning methods. Traditional methods typically use a pseudo-labeling from the scribbles using a graph method [13], [14], for example, or a regularized loss [15], [16]. The fact is that in both cases, the problems related to backpropagation continue to impact those models. On the other hand, the FLIM learning process is direct and does not require a backpropagation algorithm, taking the expert's knowledge into account.

The FLIM learning process is illustrated in Figure 1: It starts with an image with two color patterns, then the user draws markers on those regions and extracts patches centered on each marked pixel (voxel) – since each patch will have a high similarity (inner product) with itself the patch is a candidate to a convolutional filter. Then using a grouping algorithm (e.g., K-means), we can take the vector that points to the group center as a convolutional filter, using the same process to learn

This work derives from Matheus's M.Sc. Dissertation from University of Campinas, available at https://www.repositorio.unicamp.br/acervo/detalhe/1267345

convolutional weights for the following layers.

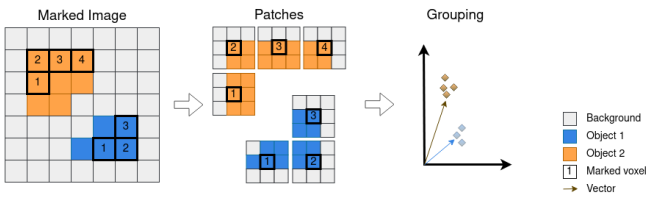

Fig. 1. FLIM learning process.

FLIM has attributes desired for a computed-aid healthcare system: it allows the user to indicate the relevant regions for filter estimation. It provides networks with fewer layers learned using user knowledge in an explanatory pipeline. It has been employed in multiple 2d natural image classification tasks [11], [17] and for detecting COVID-19 from chest CT images [18]. For segmentation, FLIM was used to assist a graph-based segmentation method [12] and to expand the segmentation in aerial images [19].

However, the literature does not provide an off-the-shelf solution nor a solution to ensure features that activate from all relevant patterns. When building a FLIM encoder for brain tumor segmentation, there might be techniques that ensure that the model will have features from markers of images with small tumors or tumors with a subtle appearance.

Figure 2 presents a few examples of those different visual appearances. It shows two pairs of MRI scans, with the FLAIR (Fluid-attenuated inversion recovery) mainly showing the edema (ED) as an active part. The second, T1Gd (T1 post-contrast with gadolinium), shows the enhancing tumor (ET) as an active part and the necrotic core (NC) as a non-active region.

It is realistic to suppose that regions with more excellent contrast, like the first pair of the left, will be more marked. In contrast, regions with small tumors and tumors with poor contrast will not be marked enough, leading to an imbalance among markers of different visual aspects. One solution is to create specific features for those images. However, with the current state of FLIM is impossible to merge other models simply due to the necessary marker-based normalization step.

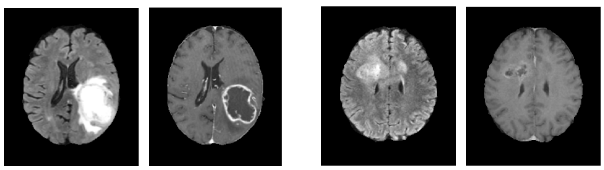

Fig. 2. Example of variability in tumor appearance. It shows two pairs of images (FLAIR and T1Gd), with the first pair containing an enormous and very expressive tumor and the second with a small and discrete.

The work's main contribution relies upon creating a complete pipeline for preprocessing and segmenting brain tumors. From the segmentation pipeline, we have the user as an active part of the learning process, drawing image markers that will be used to extract filter candidates (image patches) and using a group algorithm to reduce those candidates into convolutional filters. Then the user selects filters from the extractor based on their visual activation and updates the model with those selected filters. It is worth mentioning that this is not an iterative segmentation but a method of creating a model in multiple cycles, estimating and selecting distinct features.

This paper is organized as follows: It starts with our detailed methodology, describing the changes in the preprocessing pipeline, discussing the guideline for drawing markers and features selection, and finally describing the FLIM improvements (a complete theoretical background can be found in the second chapter of the M.Sc Dissertations [20]). After that, we describe all experiments, the adopted architecture, evaluation metrics, baseline, and SOTA comparison. Finally, we present the respective results and the work conclusion.

II. Methodology

We have created two multi-cycle learning approaches using FLIM in the form of multiple encoders, one for each MRI sequence (FLAIR and T1Gd). The first MS-FLIM joins multiple feature vectors by varying intermediate clustering hyperparameters while maintaining the same image and markers set. In the second, B-FLIM, we fixed such hyperparameters and combined feature vectors learned from different sets of images and markers.

A. Guidelines for marking and image selection

To help with this marker imbalance, we use an internal guideline for the image marking and feature selection. We first separate the tumor regions into *weak* and *strong* according to their appearance, contrast, and brightness (e.g., ED_w and ED_s). So when marking, we preferred using markers on the edges of the regions of interest, initially marking the *weak* regions and avoiding marking the *strong* regions too much. Edge marking, for example, from NC to ET, helps normalize the markers and provides approximate candidates for both patterns, in this case, NC and ET.

This division of weak/strong was also used in selecting features, where we preferred the filters that activated for the weak and subtle regions and then selected the filters of the regions of greater contrast/brightness.

B. MS-FLIM

MS-FLIM solves the unbalance problem of filter candidates (marked patches) by undersampling those candidates of each marker and image. By doing this, each scribble and image will supply the same amount of filter candidates, reducing the impact of extensive scribbles and images with enormous amounts of scribbles.

Figure 3 shows the process of MS-FLIM. It starts (1) with patch extraction of one marker (a connected scribble), which will be reduced (2) to a small number of $\mathcal{K}_m \in \mathbb{N}^*$ samples ($\mathcal{K}_m = 8$ in this example), then repeat the same process to all markers of this image (3). Then all candidates of this image will be reduced (4) to $\mathcal{K}_i \in \mathbb{N}^*$ candidates ($\mathcal{K}_i = 6$). By

doing this to all images, the result is a better environment when compared with the case without intermediates undersampling or we uses high values for $\mathcal{K}_m$ and $\mathcal{K}_i$. Finally, after the undersampling is repeated for all images, the user can carry out a third and final reduction to a desired number of filters (C_o) or, as in the case of this work, manually select the filters (i.e., candidates of all images) based on their activations.

To address the multiple features, we execute the learning process numerous times, using different values of $\mathcal{K}_m$ and $\mathcal{K}_i$. The multi-step procedure lies in the variation of $(\mathcal{K}_m, \mathcal{K}_i)$ under user control to explore different configurations of the patch feature space, extracting filters that are more specific and later filters that are more general. Both undersampling operations use K-means.

Worth mentioning that in multiple cycles of filter estimation and selection using MS-FLIM, only the hyperparameters $\mathcal{K}_i$ and $\mathcal{K}_m$ are changed, with the markers being unchanged during the whole process.

C. B-FLIM

FLIM relies on the fact that the convolution between an image patch P and a filter F, $P * F$, can be seen as the vector product between their vectorized forms:

$$P * F = \langle vec(P)_{n\times 1}, vec(F)_{n\times 1} \rangle. \quad (1)$$

However, for a better separation of candidate filters (patches), normalization of such candidates is performed, the operation called *marker-based normalization* [11], [12]. Therefore, the mean μ and the σ deviation are obtained for the set of markers $\mathcal{M}$ so that equation 1 becomes

$$P * F = \langle \sigma^{-1}(vec(P_v) - \mu), vec(F_v) \rangle \quad (2)$$

Therefore the FLIM blocks are composed of marker-based normalization followed by convolution, ReLU, and Pooling. However, even though the normalization of markers is relevant, it inhibits the joining of a model apprehended from a dataset $\mathfrak{D}_1 = (\mathcal{I}_1, \mathcal{M}_1)$ with a dataset $\mathfrak{D}_2 = (\mathcal{I}_2, \mathcal{M}_2)$, here $\mathcal{I}$ and $\mathcal{M}$ are images and markers sets.

In this work, we performed a new mathematical formulation by modifying the marker-based normalization for bias and using the convolution with the normalized filter. Then, equation 2 can be rewritten as

$$P * F \quad = \quad \langle vec(P_v), vec(F'_v) \rangle + \beta, \quad (3)$$

where $vec(F'_v) = \sigma^{-1} vec(F_v)$ and $\beta = -\mu vec(F'_v)$ is the bias.

In this way, we can merge different FLIM models (obtained from $\mathfrak{D}_1$ and $\mathfrak{D}_2$), by concatenating different normalized filters F'_1 and F'_2, as well as the bias terms β_1 e β_2, producing features that are naturally concatenated as F_{new}. Thus, the convolution between the image and this new filter is done directly using matrix multiplication not requiring the creation of parallel models or the use of multiple marker-based normalization.

D. Preprocessing

We used a modified pipeline for preprocessing brain tumors; firstly, we took the general pipeline from [21], using the registration in two steps and not using N4 to FLAIR or T1Gd. Then we added the MSP alignment based on estimating three points using a multi-scale search [22]. Also, we use a median filter and an affine registration, which have more degrees of freedom than the rigid.

We adopted an atlas-based segmentation method [23], which requires images on the same reference space, taking the MNI as the template space [24].

For the normalization we adopted an strategy based on removing the highest 1% intensities, followed by an normalization of the brain voxels.

III. Experiments

A. Qualitative results for preprocessing

We initially compared the impact of changing the preprocessing pipeline, comparing our final results with the baseline method of [21], [25], [26], particularly by using the Cancer Imaging Phenomics Toolkit (CaPTk) implementation [1]. Worth to mention that we added the MSP alignment and changed the registration template and the skull-striping methods.

B. Experimental Setup

1) Adopted architecture and its variants: The sU-Net architecture (Figure 4) consists of two encoders, one for T1Gd and the other for FLAIR images, with three convolutional layers each. Skip connections concatenate the output feature blocks, before each strided pooling operation, for both the T1Gd and FLAIR encoders, and in the final layer, a convolution with kernels 1^3 generates four channels, one for the background and one for each label (ED, ET, NC).

2) Datasets: The experiments used a private dataset containing 80 3D images of GBM with two MRI scans (FLAIR and T1Gd) per patient.

3) Encoder Training: FLIM and MS-FLIM encoders initialization were performed from the same user-drawn markers on eight volumetric images from the training set, with each image having a small number of slices marked to train the FLIM (1 to 8 slices per volumetric image). Moreover, we manually select those eight volumetric pairs of images (FLAIR and T1Gd) based on their visual aspects (tumor size, appearance, and contrast). Notice that the user is part of the learning process, so using the knowledge to select images is plausible.

We used the same hyperparameters and criteria for building the encoder, using firstly $\mathcal{K}_m = 5$ and $\mathcal{K}_i = 5$, then $\mathcal{K}_m = 10$ and $\mathcal{K}_i = 50$. For the B-FLIM, we fixed the parameters as $\mathcal{K}_m = 5$ and $\mathcal{K}_i = 5$ as a fair comparison.

For comparison proposes, we used the same set of images to train both the encoders for the biased version of FLIM (B-FLIM). But with the difference that we used extra sets of markers to explore different features as it demanded.

[1] https://cbica.github.io/CaPTk/preprocessing_brats.html

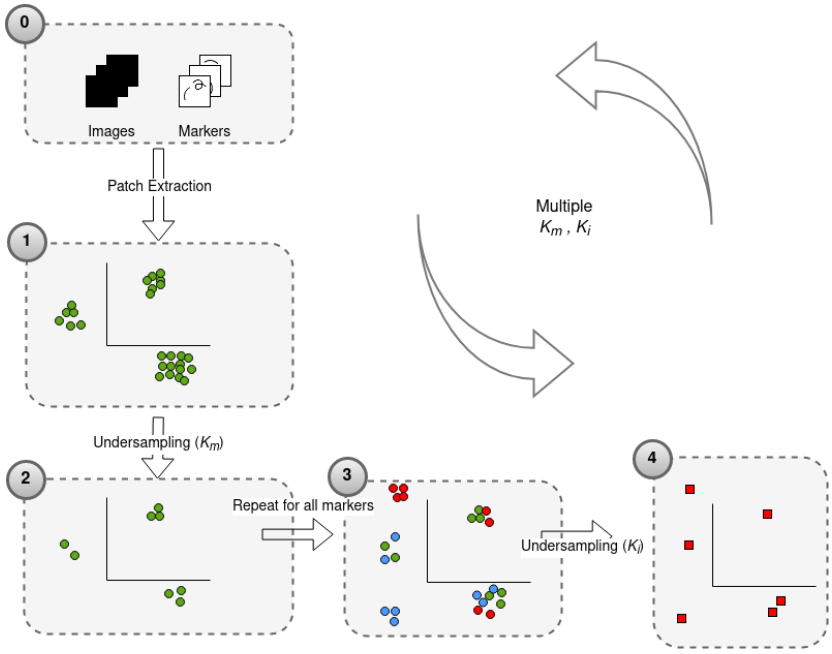

Fig. 3. MS-FLIM

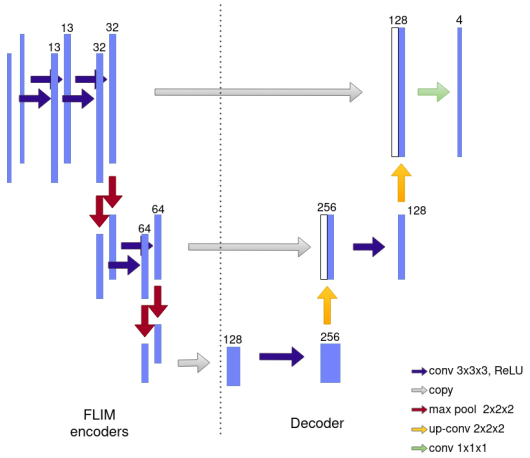

Fig. 4. sU-Net architecture with two encoders, one for FLAIR and the other for T1Gd scans.

Figure 5 presents examples of marked images used to train FLIM encoders (MS-FLIM and B-FLIM), it shows two pair of FLAIR/T1Gd samples.

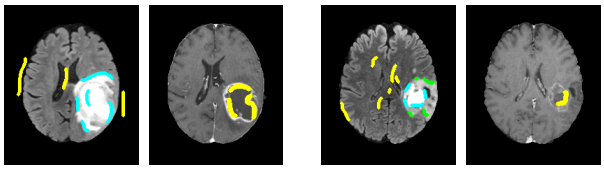

Fig. 5. Few examples of image markers used for training FLIM encoders.

4) Decoder training: Complete image annotation was used only for backpropagation, and for each run, we used the exact configuration of data split, learning rate ($2.5e^{-3}$ with linear decay), loss (average of Cross-Entropy and Dice), and a total of 100 epochs. We also used ADAM optimizer and a batch size equal to one.

C. Evaluation Metrics

We evaluate GBM tumor segmentation into three regions: ET, Tumor Core (TC) and Whole Tumor (WT). The literature usually reports the segmentation effectiveness for these three regions, assuming that $\mathrm{WT} = \mathrm{ED} \cup \mathrm{ET} \cup \mathrm{NC}$ and $\mathrm{TC} = \mathrm{ET} \cup \mathrm{NC}$. We used the Dice Similarity Coefficient (DSC) to measure efficacy and the 95-th percentile Hausdorff Distance (HD95) to verify the maximum discrepancy between the segmentation borders.

1) Golden Standard Models: DeepMedic [2] and nnU-Net [3] models were used as golden standard models. These models adopt data augmentation, normalization, and learning rate reduction, providing us with upper-bound metrics. DeepMedic is a dual-branch network that has been shown to use small amount of memory while maintaining performance [27], and nnU-Net is a very relevant network, winning segmentation challenges of the last two years [7], [8].

D. Baseline comparison

For the MS-FLIM, we compared four methods of training based on the sU-Net architecture: sU-Net and FBP, in which encoder and decoder are trained with backpropagation only (i.e., FBp - Fully backpropagation). In the sU-Net (FLIM and PBp), FLIM trains the encoder, and only the decoder is

[2]https://github.com/deepmedic/deepmedic
[3]https://github.com/MIC-DKFZ/nnUNet

trained by backpropagation (i.e., PBp - Partial Backpropagation). sU-Net (MS-FLIM and PBp), in which the encoder is trained with MS-FLIM and only the decoder is trained with backpropagation; and sU-Net (B-FLIM and PBp), with the encoder obtained from B-FLIM. Also, we verify the impact of backpropagation on the MS-FLIM (best model), verifying if FBp can improve the MS-FLIM initialization. All experiments were conducted on an Intel Xeon 2.20GHz with an NVIDIA RTX A6000 graphic card.

IV. Results

A. Preprocessing qualitative Results

Figure 6 compares our preprocessing pipeline (top) with the baseline preprocessing pipeline from BraTS (bottom) [21], [25], [26]. It shows 3 FLAIR images in which the skull-stripping of the BraTS preprocessing pipeline removes part of the tumor (pointed by red arrows). The first image (left) shows an image with removed parts of the tumor and brain regions. The second and third images show examples that the top of the brain (as well as the tumor) was removed, which can be seen by the horizontal cut aspect (red arrow). We can see how our pipeline maintained greater tumor integrity by comparing the lower and upper lines.

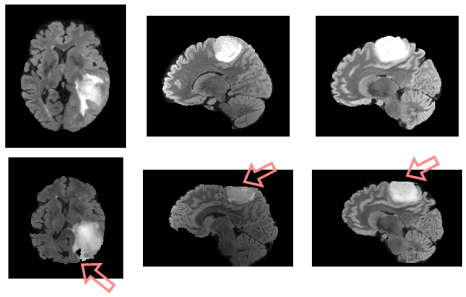

Fig. 6. Example of the quality of our pipeline regarding brain region integrity. Our pipeline (top) versus the baseline preprocessing pipeline (bottom), red arrows indicate regions where the skull-stripping procedure removes the brain tumor.

B. Our Results

Initially, we compared our methods (MS-FLIM and B-FLIM) against our baseline using sU-Net (FLIM and standard backpropagation). Therefore, Table I shows the mean and standard deviation for the DSC and HD95 metrics.

From the DSC, the MS-FLIM and the B-FLIM surpassed the standard FLIM and the backpropagation (FBp), showing the user's importance in the features' estimation and selection. Also, it shows that changes on the first layer profoundly impact the rest of the model, significantly improving the average and reducing the standard deviation of the model.

The B-FLIM provides comparable results with the MS-FLIM, even showing the best metric for the WT. However, since MS-FLIM presented the best metrics for other regions (ET/TC), we compared if the backpropagation could improve the MS-FLIM encoder (MS-FLIM with FBp versus PBP), which was different due to the similar results, a significant finding because it shows the power of training using MS-FLIM.

We also compared the results for the HD95 distance, as shown in Table I. The sU-Net trained only with backpropagation (FBp) was surpassed by the MS-FLIM and the B-FLIM. However, the standard FLIM showed the worst results, with a high mean and standard deviation for all classes. This indicates that even if a method has good dice results, it may have an erratic edge segmentation.

Figure 7 presents an example of the poor delineation from FLIM (pointed by the arrow), which impacts the HD95 metric, as shown in the table. The figure also shows the correspondent ground truth and the predictions of MS-FLIM and B-FLIM, which performed better than the standard FLIM.

Another remarkable result is that while B-FLIM provides a not-so-well HD95 metric for WT, it achieved the best DSC for the same class, illustrating the deficiency of a delineation problem of the DSC metric.

Furthermore, when comparing the general results for B-FLIM and MS-FLIM, we observed that even though the former allows generating features from a single image, the variation we obtained with MS-FLIM when varying the $\mathcal{K}_m$ and $\mathcal{K}_i$ is healthy for the model metric.

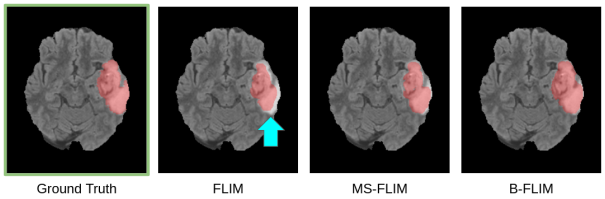

Fig. 7. Example of a lack of delineation. The figure shows a poor delineation from FLIM, which impacts the HD95 metric.

C. Comparison with SOTA

As expected, nnU-Net and DeepMedic performed better than the sU-Net-based models, as shown in the DSC of Table II. However, the differences are comparable, considering the standard deviation in the case of MS-FLIM and B-FLIM, indicating our method's high quality. This is an impressive result since models based on MS-FLIM and B-FLIM do not use massive data augmentation techniques, as done by DeepMedic and mainly by the nnU-Net.

Table II also presents our testing set's mean and standard deviation of HD95. Unexpectedly, DeepMedic achieved better results than the nnU-Net for the HD95 (We expected nnU-Net to perform better once it won the segmentation challenges).

More interesting is that MS-FLIM performed very well for both the ET and TC (having better metrics than the nnU-Net), which is surprising since MS-FLIM has slow values of DSC. This is another example of the benefits of our method and how only evaluating the DSC could be deceptive.

We also compared the cost of both models, sU-Net, nnU-Net, and DeepMedic. For that, we computed the training time

TABLE I
EVALUATION METRICS OF OUR METHOD AGAINST OUR BASELINE (STANDARD FLIM AND FBP).

Models	DSC ↑			HD95 ↓		
	ET	TC	WT	ET	TC	WT
FBp	0.665 ± 0.166	0.734 ± 0.157	0.721 ± 0.104	6.63 ± 3.45	8.04 ± 3.90	11.02 ± 2.24
FLIM and PBp	0.691 ± 0.073	0.733 ± 0.072	0.702 ± 0.109	10.39 ± 12.25	12.85 ± 13.37	16.48 ± 8.64
MS-FLIM and PBp	0.746 ± 0.052	**0.813 ± 0.073**	0.785 ± 0.082	**4.85 ± 1.85**	**6.46 ± 2.63**	10.18 ± 3.89
MS-FLIM and FBp	**0.747 ± 0.051**	0.805 ± 0.078	0.780 ± 0.096	4.93 ± 1.73	7.02 ± 2.09	**10.07 ± 2.62**
B-FLIM and PBp	0.744 ± 0.051	0.808 ± 0.077	**0.786 ± 0.085**	5.02 ± 1.88	6.83 ± 2.37	12.89 ± 8.90

TABLE II
EVALUATION METRICS OF OUR METHOD AGAINST SOTA MODELS.

Models	DSC ↑			HD95 ↓		
	ET	TC	WT	ET	TC	WT
DeepMedic	0.777 ± 0.056	0.851 ± 0.066	0.792 ± 0.094	**4.64 ± 3.15**	**7.00 ± 5.33**	9.35 ± 2,20
nnU-Net	**0.798 ± 0.045**	**0.885 ± 0.058**	**0.851 ± 0.068**	6.79 ± 8.59	7.74 ± 10.16	**7.37 ± 2.70**
sU-Net with MS-FLIM	**0.746 ± 0.052**	**0.813 ± 0.073**	0.785 ± 0.082	4.85 ± 1.85	**6.46 ± 2.63**	**10.18 ± 3.89**
sU-Net with B-FLIM	0.744 ± 0.051	0.808 ± 0.077	**0.786 ± 0.085**	**4.73 ± 1.88**	6.58 ± 2.26	11.79 ± 8.63

(TT) related to the backpropagation, the inference time (IT), and the total number of parameters (NoP), both presented in Table III.

TABLE III
BEST COST METRICS FOR EACH MODEL: TRAINING TIME (TT), INFERENCE TIME (IT), AND NUMBER OF PARAMETERS (NOP).

Metric	sU-Net (MS-FLIM with PBp)	nnU-Net	DeepMedic
TT	4789.52 s	94036.56 s	7794.90 s
IT	25.94 s	36.57 s	76.56
NoP	2455421	31197248	1060095

The results show that sU-Net has more than 12 times fewer parameters than the nnU-Net architecture, a very relevant framework for BTS. Also nnU-Net IT was computed only for the feed-forward, not considering the cropping around the brain region, a preprocessing step from then.

The motive related to DeepMedic having a small number of parameters, while it has a high number of layers, is that the layers have a limited number of outputs in each layer (i.e., 50 against 320 for the nnU-Net), which is associated with the total number of parameters. Furthermore, the high processing time is related to the fact that the network uses small patches as input against the entire image of our network.

V. CONCLUSION

In this work, we presented a complete pipeline for pre-processing and segmenting brain tumors. Our results show that our preprocessing did not lose parts of the brain or tumor compared to the baseline method. Also, from the segmentation, we presented two new approaches based on the recent FLIM methodology, MS-FLIM and B-FLIM. MS-FLIM requires multiple executions of the FLIM algorithm, using different hyperparameters to control the undersampling of patch candidates from markers and images for filter estimation. B-FLIM is more generic, allowing multiple executions of the FLIM algorithm for different hyperparameters and marker sets.

We employed both methods in the first convolutional layer, with the user selecting the most relevant filters. The results show improvements over the same architecture trained with the standard FLIM-based approach and using backpropagation only. We also showed that our methods achieve effectiveness within the standard deviation of the SOTA extensive models, even though we used a network with fewer layers.

For future works, we intend to investigate the construction of FLIM-based models with less user effort in activation observation, image, and filter selection from the training set. There is also space to combine MS-FLIM and B-FLIM, detecting the clustering hyperparameters.

VI. PUBLICATIONS

The author's dissertation yielded a conference paper related to MS-FLIM [28].

ACKNOWLEDGMENT

The authors thank FAPESP (2014/12236-1), CNPq (303808/2018-7, 407242/2021-0 and 130418/2021-8), and CAPES (88887.820768/2023-00) for financial support.

REFERENCES

[1] Q. T. Ostrom, H. Gittleman, J. Fulop, M. Liu, R. Blanda, C. Kromer, Y. Wolinsky, C. Kruchko, and J. S. Barnholtz-Sloan, "Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2008-2012," *Neuro-oncology*, vol. 17, no. suppl_4, pp. iv1–iv62, 2015.

[2] Q. T. Ostrom, M. Price, C. Neff, G. Cioffi, K. A. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2015–2019," *Neuro-oncology*, vol. 24, no. Supplement_5, pp. v1–v95, 2022.

[3] V. Simi and J. Joseph, "Segmentation of glioblastoma multiforme from mr images–a comprehensive review," *The Egyptian Journal of Radiology and Nuclear Medicine*, vol. 46, no. 4, pp. 1105–1110, 2015.

[4] C. Dupont, N. Betrouni, N. Reyns, and M. Vermandel, "On image segmentation methods applied to glioblastoma: state of art and new trends," *IRBM*, vol. 37, no. 3, pp. 131–143, 2016.

[5] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[6] Z. Jiang, C. Ding, M. Liu, and D. Tao, "Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 231–241.

[7] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnu-net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*. Springer, 2021, pp. 118–132.

[8] H. M. Luu and S.-H. Park, "Extending nn-unet for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II*. Springer, 2022, pp. 173–186.

[9] Z. Zhao, P. Xu, C. Scheidegger, and L. Ren, "Human-in-the-loop extraction of interpretable concepts in deep learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 780–790, 2021.

[10] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[11] I. E. De Souza and A. X. Falcão, "Learning cnn filters from user-drawn image markers for coconut-tree image classification," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[12] I. E. de Souza, B. C. Benato, and A. X. Falcão, "Feature learning from image markers for object delineation," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2020, pp. 116–123.

[13] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.

[14] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble-supervision alone," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 236–244.

[15] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised cnn segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 507–522.

[16] R. Dorent, S. Joutard, J. Shapey, S. Bisdas, N. Kitchen, R. Bradford, S. Saeed, M. Modat, S. Ourselin, and T. Vercauteren, "Scribble-based domain adaptation via co-segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 479–489.

[17] B. C. Benato, I. E. de Souza, F. L. Galvão, and A. X. Falcão, "Convolutional neural networks from image markers," *arXiv preprint arXiv:2012.12108*, 2020.

[18] A. M. Sousa, F. Reis, R. Zerbini, J. L. Comba, and A. X. Falcão, "Cnn filter learning from drawn markers for the detection of suggestive signs of covid-19 in ct images," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3169–3172.

[19] I. E. de Souza, C. L. Cazarin, M. R. Veronez, L. Gonzaga, and A. X. Falcão, "User-guided data expansion modeling to train deep neural networks with little supervision," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[20] M. A. Cerqueira, "User-assisted design of a neural network for brain tumor segmentation," Master's thesis, Universidade Estadual de Campinas, Instituto de Computação, 2023.

[21] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[22] G. C. Ruppert, L. Teverovskiy, C.-P. Yu, A. X. Falcao, and Y. Liu, "A new symmetry-based method for mid-sagittal plane extraction in neuroimages," in *2011 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE, 2011, pp. 285–288.

[23] S. B. Martins, J. Bragantini, A. X. Falcão, and C. L. Yasuda, "An adaptive probabilistic atlas for anomalous brain segmentation in mr images," *Medical physics*, vol. 46, no. 11, pp. 4940–4950, 2019.

[24] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. Almli, and D. Collins, "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood," *NeuroImage*, vol. 47, p. S102, 2009.

[25] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[26] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[27] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[28] M. A. Cerqueira, F. Sprenger, B. C. Teixeira, and A. X. Falcão, "Building brain tumor segmentation networks with user-assisted filter estimation and selection," in *18th International Symposium on Medical Information Processing and Analysis*, vol. 12567. SPIE, 2023, pp. 202–211.