frontiers in Radiology

ORIGINAL RESEARCH
published: 23 September 2021
doi: 10.3389/fradi.2021.711514

Meningioma Segmentation in T1-Weighted MRI Leveraging Global Context and Attention Mechanisms

David Bouget[1], André Pedersen[1], Sayied Abdol Mohieb Hosainey[2], Ole Solheim[3,4] and Ingerid Reinertsen[1]*

[1] Department of Health Research, SINTEF Digital, Trondheim, Norway, [2] Department of Neurosurgery, Bristol Royal Hospital for Children, Bristol, United Kingdom, [3] Department of Neurosurgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, [4] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway

Purpose: Meningiomas are the most common type of primary brain tumor, accounting for ∼30% of all brain tumors. A substantial number of these tumors are never surgically removed but rather monitored over time. Automatic and precise meningioma segmentation is, therefore, beneficial to enable reliable growth estimation and patient-specific treatment planning.

Methods: In this study, we propose the inclusion of attention mechanisms on top of a U-Net architecture used as backbone: (i) Attention-gated U-Net (AGUNet) and (ii) Dual Attention U-Net (DAUNet), using a three-dimensional (3D) magnetic resonance imaging (MRI) volume as input. Attention has the potential to leverage the global context and identify features' relationships across the entire volume. To limit spatial resolution degradation and loss of detail inherent to encoder–decoder architectures, we studied the impact of multi-scale input and deep supervision components. The proposed architectures are trainable end-to-end and each concept can be seamlessly disabled for ablation studies.

Results: The validation studies were performed using a five-fold cross-validation over 600 T1-weighted MRI volumes from St. Olavs Hospital, Trondheim University Hospital, Norway. Models were evaluated based on segmentation, detection, and speed performances, and results are reported patient-wise after averaging across all folds. For the best-performing architecture, an average Dice score of 81.6% was reached for an F1-score of 95.6%. With an almost perfect precision of 98%, meningiomas smaller than 3 ml were occasionally missed hence reaching an overall recall of 93%.

Conclusion: Leveraging global context from a 3D MRI volume provided the best performances, even if the native volume resolution could not be processed directly due to current GPU memory limitations. Overall, near-perfect detection was achieved for meningiomas larger than 3 ml, which is relevant for clinical use. In the future, the use of multi-scale designs and refinement networks should be further investigated. A larger number of cases with meningiomas below 3 ml might also be needed to improve the performance for the smallest tumors.

Keywords: 3D segmentation, attention, deep learning, meningioma, MRI, clinical diagnosis

OPEN ACCESS

Edited by:
Jaeil Kim,
Kyungpook National University,
South Korea

Reviewed by:
Shan Lin,
University of Washington,
United States
Christoph von Tycowicz,
Freie Universität Berlin, Germany

****Correspondence:***
David Bouget
david.bouget@sintef.no

Specialty section:
This article was submitted to Artificial Intelligence in Radiology, a section of the journal Frontiers in Radiology

Received: *18 May 2021*
Accepted: *16 August 2021*
Published: *23 September 2021*

Citation:
Bouget D, Pedersen A, Hosainey SAM, Solheim O and Reinertsen I (2021) Meningioma Segmentation in T1-Weighted MRI Leveraging Global Context and Attention Mechanisms. Front. Radiol. 1:711514. doi: 10.3389/fradi.2021.711514

1. INTRODUCTION

Primary brain tumors, characterized by an uncontrolled growth and division of cells, can be grouped into two main categories: gliomas and meningiomas. Gliomas represent the highest mortality rate (1) while meningiomas account for one-third of all operated central nervous system tumors (2). The prevalence rate of meningiomas in the general population undergoing 1.5 T non-enhanced magnetic resonance imaging (MRI) scans is 0.9% (3). Recently, the increase in incidence is presumably due to higher detection rates from a widespread use of MRI in the general population (4). Many meningiomas are encountered as incidental findings on neuroimaging, but never surgically removed. According to the EANO consensus guidelines (5), asymptomatic patients can be managed through observation only. An annual MRI follow-up of benign meningiomas (i.e., WHO grade I) is recommended, then biennal after 5 years. Surgery is then usually indicated if a follow-up shows tumor growth. Today, growth assessment in a clinical setting is routinely based on eyeballing or crude measures of tumor diameters (6). Manual segmentation by radiologists is time consuming, tedious, and subject to intra-/inter-rater variations difficult to characterize (7) and is therefore rarely done in clinical routine. Systematic and consistent brain tumor segmentation and measurements through (semi-)automatic methods are consequently of utmost importance. From accurate tumor growth measurement and future growth estimation, patient-specific follow-up plans could potentially be enabled. Moreover, assessing the growth pattern on individual level may be informative with respect to treatment indication, as a majority may exhibit a self-limiting growth pattern (8). Finally, segmentation is key for assessing treatment response after radiotherapy or surgery. For medical imaging, MRI represents the gold standard due to its non-invasiveness and widespread availability. A contrast-enhanced T1-weighted sequence is most often favored, rendering the tumor border more easily distinguishable (5). Alternatively, the fluid-attenuated inversion recovery (FLAIR) sequence can complement the diagnosis. Effects of fluids on the image are nullified, enabling a better visualization of edema regions. Nevertheless, inherent downsides can be associated with MRI acquisitions such as intensity inhomogeneity (9), variations from the use of different acquisition scanners (10), or variations in the acquisitions (e.g., field-of-view, slice thickness, or resolution). In T1-weighted MRI, meningiomas are often sharply circumscribed with a strong contrast enhancement, making them clear to identify. However, small meningiomas might resemble other contrast-enhanced structures such as blood vessels, hindering the detection task. In order to alleviate radiologists' burden to annotate large contrast-enhanced meningiomas, while at the same time to help detecting smaller and unusual meningiomas, automatic segmentation methods are paramount.

In recent years, automatic and end-to-end semantic segmentation has known considerable improvements through the development of fully convolutional neural network architectures (FCNs) (11–13). By restoring the feature map of the last deconvolution layer to the size of the initial input sample, predictions can be generated for each voxel. While such architectures provide near radiologist-level performances on some medical image analysis tasks (14, 15), multi-stage cascading induces a loss of local information leading to excessive and redundant low-level features. The most effective solution to boost the segmentation performance is to combine local and global information to preserve consistency in the feature maps. However, 3D medical volumes are typically too sizable to fit on GPU memory at their original resolution, and the number of parameters for the corresponding model would be considerable. Different trade-offs have been investigated such as splitting the 3D volume into a series of patches or slabs by which some global context can be leveraged while good local information is retained (16). Capturing the entire global context from a full 3D volume is important for a model to understand the spatial relationships between the different anatomical structures. Aggregating multi-scale contexts and using various dilated convolutions and pooling operations can be a solution (17, 18). Instead, capturing richer global information through enlarged kernels (19) or fusing semantic features at different levels (20) can cope with information loss but are unable to leverage overall relationships between structures. To address shortcomings from feature maps consistency and loss of information when using multi-stage cascading architectures, attention mechanisms have been utilized with great success (21–23). Attention modules can be seamlessly coupled with regular FCN architectures for end-to-end training, with the benefit of letting the model learn to focus on the segmentation targets without significant computational overhead. Optimally coupled with each deconvolution block, attention can be designed to capture features' dependencies spatially, channel-wise, or across any other dimension (21). Alternatively, multiple models operating on different input shapes or focusing on different aspects during training can be fused as a post-processing step, called an ensemble, to generate the final prediction map (24). Global context and local refinement can virtually be obtained separately at the cost of longer training and inference time, and higher model complexity. However, ensembling has not always shown to produce better overall segmentation performance compared to a single model's use (25).

The Multimodal Brain Tumor Image Segmentation (BraTS) challenge dataset represents a cornerstone in the field of brain tumor segmentation. Featuring only patients with high-/low-grade gliomas, it fostered the development of many methods in the community (26). At first, and due to memory limitations, the task of brain tumor segmentation has been approached in 2D where each axial image (i.e., slice) from the original 3D MRI volume was processed sequentially. Havaei et al. hinted at the benefits from combining the immediate local neighborhood and a larger context such as the overall position in the brain. Local and global information were fused within a two-pathway convolutional neural network (CNN) with multi-modal inputs (27). Recurrent neural networks, using image patches and slices along the three different acquisition planes (i.e., axial, coronal, and sagittal), were alternatively investigated (28). The predictions from the different CNNs were fused using a

voting-based strategy. Other methods relying on image or patch-based strategies have also been proposed to deal with large MRI volumes in an efficient way (29–31). Features obtained from image patches or through a slabbing process (i.e., using a set of slices) will inherently contain limited global information. As such, methods based upon these conducts will generally achieve lower performance than methods leveraging features extracted from the entire 3D volume. Simple 3D CNN architectures (32, 33), multi-scale approaches (34, 35), and ensembling of multiple CNNs (24) have hence been explored. Better segmentation performance, increased robustness toward hyperparameters, and improved capability to generalize were exhibited. However, the stacking strategy inherent to ensembling leads to longer and more cumbersome procedures for training and inference. The potential from efficiently computing meaningful features from a whole 3D MRI volume remains yet to be fully explored.

Outside the scope of the BraTS challenge dataset, the meningioma segmentation task has been scarcely investigated. A multi-modal (T1c, T2f) and multi-class (core tumor and edema) segmentation has been attempted using traditional machine learning methods (e.g., SVM) (7). Unfortunately, the validation studies have been carried out on a dataset of only 15 patients, making it difficult to fully assess the ability to generalize. The DeepMedic architecture and framework (34), operating patch-wise in 3D, has been investigated by Laukamp et al. on their own multi-modal dataset (36, 37). A combination of T1-weighted contrast-enhanced and FLAIR sequences was used as input for the segmentation of contrast-enhancing tumor volume and total lesion including surrounding edema. The limited validation group of 56 patients and the need for a second 3D fully connected network in post-processing to remove false positives were legitimate disadvantages. In our previous work, leveraging a whole MRI volume, rather than slab-wise, has shown to boost the overall segmentation performance (38). However, using a regular 3D U-Net or multi-scale architecture still resulted in the loss of information in the encoding path which remained to be addressed. To summarize, recurring identified limitations from previous meningioma segmentation studies include the relatively minimal datasets used with at most 126 patients for an average meningioma volume of 31.5 ml. In addition, the lack of advanced validation studies to prove generalization, and the common pitfalls from slab/patch-wise methods or inefficient architectures for capturing large-scale relationships were identified. While relying on multiple modalities as input is of interest, the sole use of T1-weighted MRI presents the benefit of being the bare minimum to open for clinical use in a screening context or at the outpatient clinic.

In this paper, we focus on reducing the information loss for encoder–decoder architectures using combinations of attention, multi-scale, and deep supervision schemes. In addition, we chose to rely on T1-weighted MRI volumes only as input to maximize the potential for use in clinical settings. Our contributions are as follows: (i) the investigation of architectures able to better understand global context, (ii) validation studies focusing on meningioma volumes for clinical and diagnostic use, and (iii) online availability for our trained models along with the inference script.

2. METHODS

2.1. Related Work

Typically, semantic features are extracted along the encoding path for encoder–decoder architectures. The field-of-view is progressively enlarged via strided convolutions or pooling operations, hence provoking some loss of detail. In the decoding path, extracted features are exploited to solve the task at hand (i.e., classification, segmentation). At the end of the encoding path, the feature maps are the richest in global relationships. Yet, limited spatial details are preserved due to cascaded convolutions and nonlinearities. In order to recover fine-grained details, symmetrical architectures (e.g., U-Net) propagate feature maps across corresponding encoder and decoder at the same level, also known as long skip connections. In general, efficient architectures optimally use global and contextual information from high-level features and border information from low-level features to resolve small details (39). Attention mechanisms focus on identifying salient image regions to amplify their influence. By filtering away irrelevant and potentially confusing information from other regions, the prediction become more contextualized (40). Hard attention, stochastic and non-differentiable, relies on sampling-based training making optimizing models more difficult. Soft attention, probabilistic and amenable to training by backpropagation, can be by contrast seamlessly integrated into current CNN architectures. Numerous tasks have benefited from attention, such as text understanding and semantic segmentation (41–43).

In a main body of work, a single attention gating is performed at every level along the decoding path. Attention feature maps are often concatenated with the feature maps from the long skip connection (22, 44). Nevertheless, propagation of the lowest-level feature maps in an upward fashion with short skip connection has also been investigated (45). In a second body of work, authors have investigated the computation of specific attention feature maps to focus on position, channel, or class dependencies. Fu et al. (21) presented a dual attention network for scene segmentation where position and channel attention modules were computed at the bottom of a ResNet encoding path. The generation of the final probability map, right after and without a matching decoding path, is detrimental to the spatial segmentation quality. Following the same idea, Mou et al. (46) added a complete ResNet decoding path after position and channel attention computation, improving the spatial reconstruction. Attempts have been made to include dual attention modules at every stage of a ResNet architecture, either from the skip connection feature maps from the encoder path (23), or in the decoder path after concatenation with the feature maps from the previous level (47). To deal with the substantial number of parameters and prevent training hurdles (e.g., overfitting, slow convergence), additional steps are required. The use of dilated convolutions, or the addition of a significant dropout over the attention feature maps, has been proposed. Finally, other hybrid attention schemes have been explored, for example in the context of aerial image segmentation with concepts such as class channel attention to

exploit dependencies between classes and generate class affinity maps (48).

To compensate for the loss of detail inherent to consecutive pooling operations, new architecture designs or layers have been proposed. In order to preserve details in the encoding path, various multi-scale attempts have been made, such as infusing down-sampled version of the input volume in each encoder block (44). Alternatively, the receptive fields can be enlarged using atrous convolutions and pyramid spatial pooling (17, 39). Lastly, the feature maps from each encoder block can be concatenated, and the created multi-scale feature maps used for guiding in an upward skip connection fashion (47). In the latter case, complementary low-level information and high-level semantics are encoded jointly in a more powerful representation. Conversely, intermediate feature maps generated at each level of an encoder–decoder architecture can be leveraged instead of computing the loss simply from the last decoder step, commonly referred to as deep supervision (DS). The rationale is that the feature maps from hidden layers of a deep network can serve as a proxy to improve the overall segmentation quality and sensitivity of the model, while alleviating the problem of vanishing gradients (49). The final loss is computed as a weighted average between the losses from each level whereby each can contribute equally (44), or with weights defined as trainable parameters. Intermediate losses can be computed separately from the raw feature maps and the attention feature maps, before tallying the final loss across all levels (47). In general, the combination of multi-resolution and deep supervision has shown to improve convergence (i.e., better optimum and faster solving) for inverse problems (50).

2.2. Dataset

In a previous study (38), we introduced a dataset of 698 Gd-enhanced T1-weighted MRI volumes acquired on 1.5 T and 3T scanners in the catchment region of the Department of Neurosurgery at St. Olavs hospital, Trondheim University Hospital, Norway. In this study, we kept the 600 high-resolution MRI volumes having a maximum spacing of 2 mm , leaving aside the remaining 98 volumes. Of those 600 patients, 276 underwent surgery to resect the meningioma, while the remaining 324 were followed at the outpatient clinic. In the dataset, MRI volume dimensions covered $[240; 512] \times [224; 512] \times [18; 290]$ voxels and the voxel sizes ranged between $[0.47; 1.05] \times [0.47; 1.05] \times [0.60; 2.00]\,\text{mm}^3$. The volumes of the surgically resected meningiomas were on average larger (30.92 ± 33.10 ml), compared to the untreated meningiomas followed at the outpatient clinic (7.62 ± 13.67 ml). Overall, meningioma volumes ranged between [0.07, 167.99] ml for an average value of 18.33 ± 27.20 ml.

2.3. Architecture Design

In this work, we opted for a U-Net architecture as backbone, which we set to five levels and used filter sizes of [16, 32, 128, 256, 256] for each level, respectively. Our first proposed architecture, named AGUNet and illustrated in **Figure 1**, integrates an attention-gated mechanism to U-Net. Our second architecture, named DAUNet and illustrated in **Figure 2**, integrates a dual attention module to U-Net. In addition, both architectures are combining multi-scale input and deep supervision support. For viewing purposes and clarity, we chose to display our proposed architectures with only three levels. The proposed design is modular whereby the backbone can be changed (e.g., U-Net, ResNet) and each main module (i.e., multi-scale input, attention mechanism, and deep supervision) can be toggled. Such design enables seamless end-to-end training while providing unbiased and comparable results. The specifics of each module are presented in the following subsections.

2.3.1. Attention Mechanisms

In our first architecture, attention gates were incorporated to each decoder step of the architecture to highlight salient features passing through the skip connections, as described previously (22). Attention gating is performed before the concatenation operation in order to merge only relevant activations. Performing gating from features extracted at a coarser scale allows for the disambiguation of irrelevant responses in skip connections. At each decoder level, the feature maps from the previous level (i.e., coarser scale) are first resampled to match the shape of the skip connection feature maps, using a transpose convolution operation with a $3 \times 3 \times 3$ kernel size (cf. green block in **Figure 1**). Inside the attention gate (cf. yellow block in **Figure 1**), the upsampled feature maps (denoted as g) and the feature maps from the skip connection (denoted as x^l) are processed to generate the gating signal. Then, the signal is applied to x^l in order to generate the gated feature maps for the current level $\hat{x}^l$. Linear transformations without spatial support (i.e., $1 \times 1 \times 1$ convolutions) are performed to limit the computational complexity and number of trainable parameters, similarly to non-local blocks (51). We chose to include an attention gate on the lowest-level feature maps (i.e., first skip connection) even though limited benefits are expected since the input data tend to not be represented in a high-enough dimensional space (40).

As second architecture, a dual attention scheme with position and channel attention modules was integrated to the U-Net architecture. Due to GPU-memory limitations and to reduce the computational complexity, the attention feature maps are only computed once at the end of the encoding path rather than at every decoder level. The position attention module, or spatial attention module, encodes a wider range of contextual information into local features to enhance their representation capability. The channel attention module exploits inter-dependencies between channel maps to emphasize inter-dependent feature maps and improve the feature representation of specific semantics, as presented by Fu et al. (21). For training efficiency, a spatial dropout operation with a rate of 0.5 and linear transformations are performed on the raw attention feature maps, generating the final Attention Feature Maps (AFMs). In a variant, named DAGUNet, the attention feature maps are propagated upward and concatenated at each decoder level to guide the computation of feature maps at the higher levels (cf. green arrow in **Figure 2**). Transferring the bottom attention feature maps requires less trainable parameters overall than

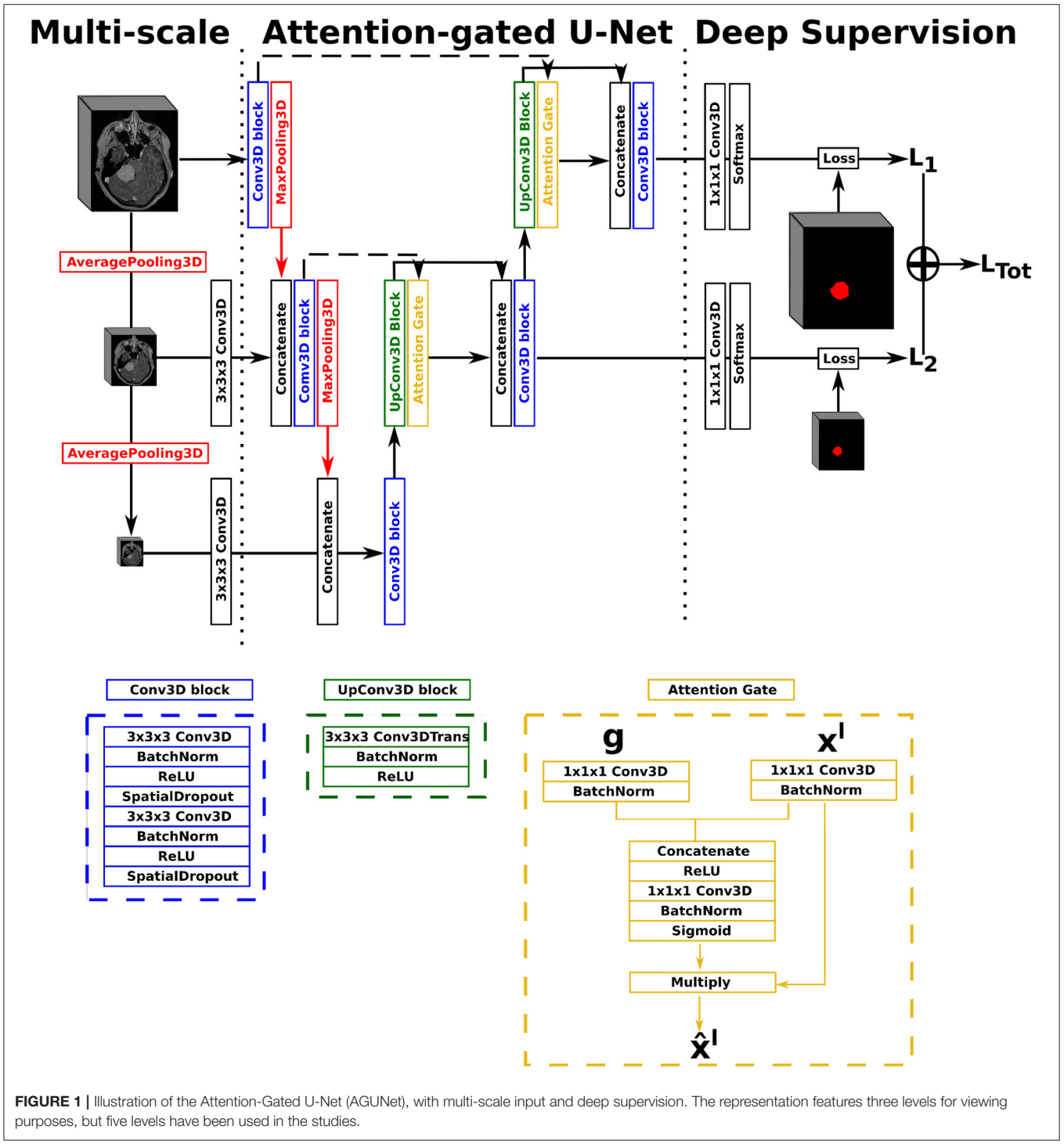

FIGURE 1 | Illustration of the Attention-Gated U-Net (AGUNet), with multi-scale input and deep supervision. The representation features three levels for viewing purposes, but five levels have been used in the studies.

computing the dual attention blocks at every decoder step, while still benefiting from them.

2.3.2. Multi-Scale and Deep Supervision

For our multi-scale approach, we opted to exploit down-sampled versions of the initial network input, at every level in the encoding path, by performing consecutive average pooling operations with a $3 \times 3 \times 3$ kernel size. Each down-sampled volume is then concatenated to the feature maps coming from the previous encoding level, before generating the feature maps for the current level, in order to preserve spatial details. For our deep supervision scheme, the ground truth volume is recursively down-sampled to match the size of the feature maps at each corresponding decoder level, where an intermediate loss L_x is computed. The final loss,

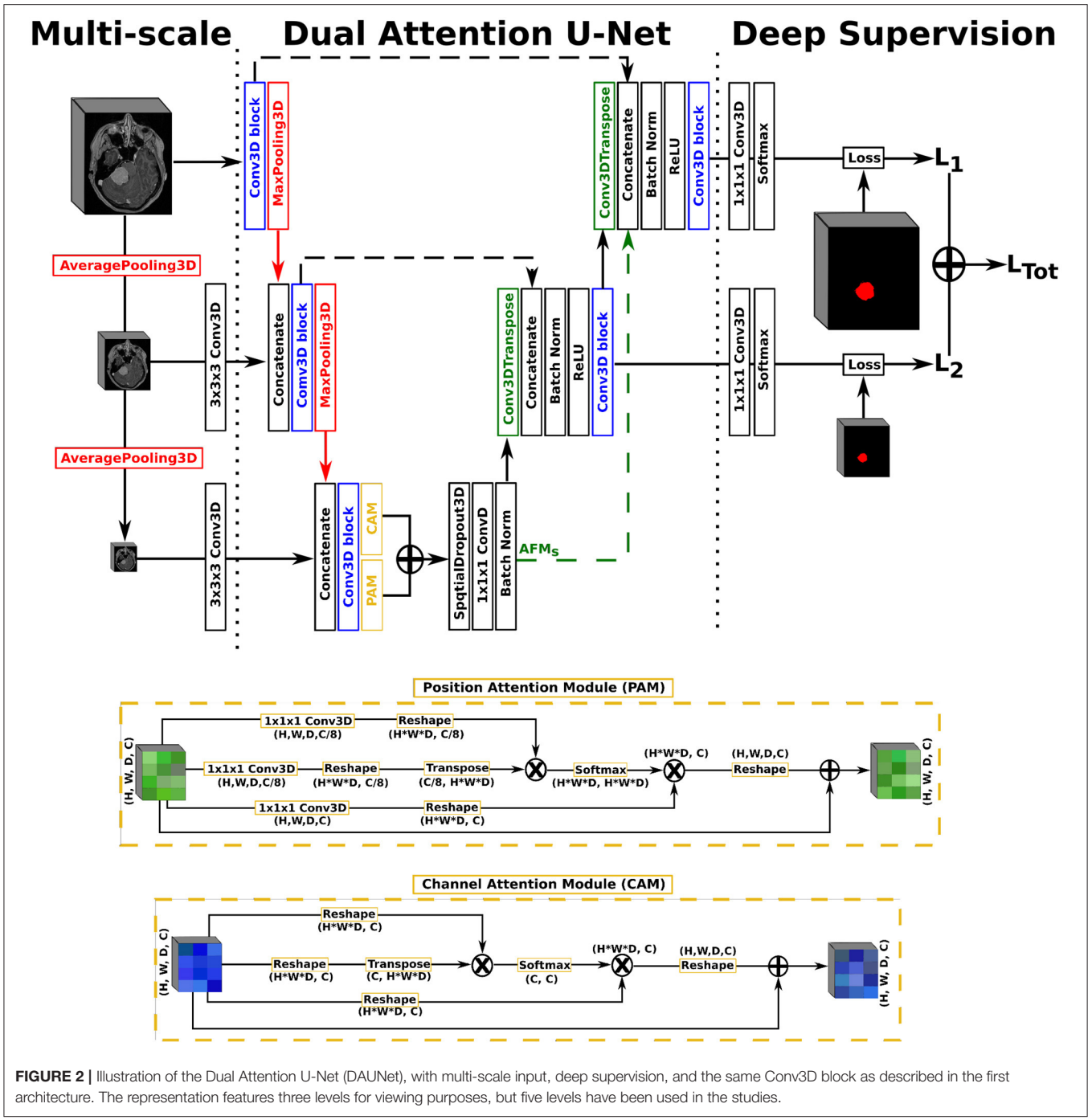

FIGURE 2 | Illustration of the Dual Attention U-Net (DAUNet), with multi-scale input, deep supervision, and the same Conv3D block as described in the first architecture. The representation features three levels for viewing purposes, but five levels have been used in the studies.

represented as L_{Tot} in **Figure 1**, is the weighted sum from all intermediate losses. In this study, we did not set the weights as trainable parameters, not to favor the feature maps from any level, and kept a uniform weighting strategy.

2.4. Training Strategies

The MRI volumes were all preprocessed using the following steps: (i) resampling to an isotropic spacing of 1 mm^3 using spline interpolation order 1 from NiBabel[1], (ii) clipping tightly around the patient's head, (iii) volume resizing to 128 × 128 × 144 voxels using spline interpolation order 1, and (iv) normalizing intensities to the range [0, 1]. A typical data augmentation approach was favored, where the following transforms were applied to each input sample with a probability of 50%: horizontal and vertical flipping, random rotation in the

[1]https://github.com/nipy/nibabel.

range $[-20, 20]^\circ$, translation up to 10% of the axis dimension, zoom between [80, 120]% in the axial plane.

All models were trained from scratch using the Adam optimizer with an initial learning rate of 10^{-3} and training was stopped after 30 consecutive epochs without validation loss improvement. The main loss function used was the class-average Dice loss, excluding the background. Additionally, we experimented with the Focal Tversky Loss (FTL), where the Tversky similarity index helps balance false positive and false negative predictions more flexibly. The focal aspect increases the contribution of hard training examples in the loss computation (44). We used $\alpha = 0.7$ and $\beta = 0.3$ for the Tversky index to minimize false negative predictions, and $\gamma = 2.0$ as focal parameter. Unless specified otherwise, all models were saved based on the minimum overall validation loss, which corresponds to L_{Tot} if deep supervision is enabled.

Given the sizable memory footprint, all models were trained using two samples in a batch. In order to improve generalization, we used the concept of accumulated gradients to effectively increase the batch size. Mini-batches up to 32 elements have shown to produce better models (52). For a specified number of accumulated gradient steps (n), each batch is run sequentially using the same model weights for calculating the gradients. When the n steps are performed, the accumulated sum of gradients from each step amounts to the same gradients as if computed over the larger batch size, ensuring that the model weights are properly updated. For our studies, we chose to perform 16 steps, enabling us to use a batch size of 32.

3. VALIDATION STUDIES

In this work, we focus primarily on optimizing segmentation and detection performance. In parallel, runtime capabilities, potential for diagnostic purposes, and clinical use are investigated. A five-fold cross-validation approach was followed whereby at every iteration three-folds were used for training, one for validation, and one for testing. Each fold was populated in order to exhibit a similar meningioma volume distribution, as described in our previous study (38).

For quantifying the performance, we used: (i) the Dice score, (ii) the F1-score, and (iii) the training/inference speed. The Dice score is used to assess the quality of the segmentation pixel-wise, between the manual ground truth (GT) and the output of the trained model (Pred), and is reported in %. The F1-score assesses the harmonic average of recall and precision instance-wise, and is reported in %. In the case of multifocal meningiomas, each foci is considered as a separate instance. Finally, the training speed is reported in hours, while the inference speed and the total processing speed to generate results for a new MRI volume are reported in seconds. For the segmentation task, the Dice score is computed between the ground truth and a binary representation of the probability map generated by a trained model. The binary representation is computed for ten different equally spaced probability thresholds (PT), in the range $[0, 1]$. A connected components approach, coupled to a pairing strategy, was employed to compute the recall and precision values. Such step is mandatory for the minority of multifocal meningiomas, but also to separate the correct prediction over a meningioma from the false positives per patient, enabling to also report the true positive Dice (Dice-TP). Pooled estimates, computed from each fold's results, are computed for each measurement (53), and reported with mean and standard deviation.

3.1. Ablation Study

Comparison of segmentation performances using various combinations of the methodological components introduced in section 2. The name given to each experiment is a concatenation of components' abbreviations. The architectures to choose from are as follows: regular U-Net (UNet), attention-gated U-Net (AGUNet), dual attention U-Net (DAUNet), and dual attention guided U-Net (DAGUNet), combined with multi-scale input (MS), deep supervision (DS), and the use of accumulated gradients (AG). If not specified otherwise, the Dice loss function is used and the best model is selected based on the total loss L_{Tot}. Usage of the focal Tversky loss is indicated by the TFL tag, while saving the best model based on the loss from the upper level is indicated by the Top tag.

3.2. Segmentation and Detection Performances Study

A comparison is performed between the best trained model for each of the main designs: slab-wise U-Net (UNet-Slabs) and PLS-Net studied previously (38), full volume U-Net (UNet-FV) and the best method identified in the ablation study (Ours). All models were compared using the exact same methodology considering only the probability threshold PT, without any consideration toward the absolute size or relative overlap of the meningioma candidates. In addition, the nnU-Net approach (54) has been selected to serve as external baseline. The optimal preprocessing steps were determined automatically from the dataset using the nnU-Net framework. A spacing of $0.93 \times 0.93 \times 1.0$ mm was selected for a median MRI volume resolution of $267 \times 265 \times 162$ voxels, leading to the selection of the 3D full resolution U-Net configuration. All models were trained for 1,000 epochs, using the joint Dice and cross-entropy loss function, operating over patches of $160 \times 160 \times 96$ voxels. The model prediction heatmaps were taken before post-processing.

3.3. Volume-Based Performances Analysis

To study the potential for clinical use in the hospital or outpatient clinic, performances are analyzed over different meningioma groups based on volume. Limitations such as challenging meningiomas and potential outliers are also described.

3.4. Speed Performances Study

For the different experiments considered in the first two validation studies, a deeper analysis around speed is conducted. The model complexity as total number of parameters and the training behavior as $s.epoch^{-1}$ (in seconds), best epoch, and total training time (in hours), are first considered. The pure inference speed is reported when using GPU and CPU (in s). Finally, the total elapsed time required to generate predictions for a new patient is reported as processing time (in s), obtained with GPU

support. The operations required to prepare the data to be sent through the network, to initialize the environment, to load the trained model, and to reconstruct the probability map in the referential space of the original volume are accounted for. The experiment has been repeated 10 consecutive times over the same MRI volume for each model, using a representative sample of $256 \times 256 \times 192$ voxels with $1.0 \times 1.0 \times 1.0$ mm spacing.

4. RESULTS

Models were trained across different machines using either an NVIDIA Quadro P5000 (16 GB) or a Tesla P100 PCIe (16 GB) dedicated GPU and regular hard drives. For inference and processing speed computation, an Intel Xeon @3.70 GHz (6 cores) CPU and an NVIDIA Quadro P5000 GPU were used. Implementation was done in Python 3.6 using `TensorFlow` v1.13.1, Cuda 10.0, and the Imgaug Python library for the data augmentation methods (55). Due to randomness during weight initialization and optimization, a fixed seed was set to make comparisons between experiments fair and reproducible. Trained models and inference code are made publically available at https://github.com/dbouget/mri_brain_tumor_segmentation.

4.1. Ablation Study

Pixel-wise segmentation and patient-wise detection performances for the different architectural designs considered are summarized in **Table 1**. The first row provides baseline results using the backbone architecture only. The greatest impact comes from the deep supervision component increasing the Dice score by about 5% and the F1-score by 2.5% between experiments (ii) and (iii). From the multi-scale input approach, $<$1% improvement for the same metrics is reported, as can be seen between experiments (iii) and (iv). It is worth mentioning that models trained using deep supervision produce comparable results whether saved based on the best total loss or the best loss from the upper level only [cf. experiments (v) and (vi)]. The use of attention modules does not further improve the results [cf. experiments (i) and (ii)]. Similarly, no added value has been recorded when using a more complex dual attention scheme [cf. experiments (v) and (x)]. A similar conclusion can be drawn for the use of the accumulated gradients strategy, degrading slightly the overall segmentation performances, for a reduction in standard deviation across detection results [cf. experiments (iv) and (v)]. While the implementation seems correct, identifying the best batch size is difficult and heavily dependant on the dataset size and diversity. However, for our second architecture with dual attention, propagating the attention feature maps upward seems to be beneficial. A increase of 1–2% across the different measurements is reported when compared to no propagation [cf. experiments (viii) and (x)]. The attempt to use the Focal Tversky loss was not conclusive as all metrics score lower in experiment (vii) compared to experiment (v).

Overall, we consider the best-performing model to be obtained by experiment (iv), reaching the highest scores for all but one metric. As we do favor detection performances over pixel-wise segmentation accuracy, our AGUNet-MS-DS model is also reaching the highest F1-score with 95.58%. In the rest of the paper, we refer to AGUNet-MS-DS [experiment (iv)] as ours.

4.2. Segmentation and Detection Performances Study

For the four different training concepts considered, segmentation performances have been reported in **Table 2**. The UNet-Slabs approach yields surprisingly competitive recall performances with only a 2% shortfall compared to our best-performing method. However, the generation of a larger amount of false positives per patient is an inherent limitation of slabbing 3D volume. The 20% difference in precision between the same two approaches is a clear testimony. While the PLS-Net architecture drastically increases the precision from leveraging a full 3D volume, its shallow architecture is not able to compete in terms of overall pixel-wise segmentation or recall performances. Nevertheless, it indicates how well global spatial relationships can be modeled and how beneficial it can be for a 3D segmentation

TABLE 1 | Performances obtained by component ablation, averaged over the five-folds. The components are as follows: regular U-Net (UNet), attention-gated U-Net (AGUNet), dual attention U-Net (DAUNet), dual attention guided U-Net (DAGUNet), multi-scale input (MS), deep supervision (DS), and the use of accumulated gradients (AG).

Experiment	PT	Dice	Dice-TP	F1	Recall	Precision
(i) UNet-FV	0.5	76.91 ± 28.98	84.77 ± 16.22	93.19 ± 01.70	90.70 ± 01.90	95.86 ± 02.28
(ii) AGUNet-AG	0.4	75.14 ± 30.38	84.21 ± 16.57	92.15 ± 01.74	89.21 ± 02.38	95.35 ± 02.37
(iii) AGUNet-DS-AG	0.4	80.72 ± 24.98	86.79 ± 12.19	94.73 ± 00.76	93.02 ± 02.01	96.63 ± 02.95
(iv) AGUNet-MS-DS	0.4	**81.64 ± 25.33**	**87.69 ± 12.12**	**95.58 ± 02.24**	93.03 ± 04.13	**98.39 ± 01.43**
(v) AGUNet-MS-DS-AG	0.4	79.49 ± 26.38	87.02 ± 11.59	94.23 ± 00.88	91.69 ± 01.73	96.93 ± 01.04
(vi) AGUNet-MS-DS-AG-Top	0.5	79.89 ± 26.52	86.64 ± 13.75	94.53 ± 08.23	92.19 ± 02.21	97.07 ± 01.67
(vii) AGUNet-MS-DS-AG-FTL	0.7	74.27 ± 30.29	84.21 ± 14.47	91.27 ± 01.50	88.20 ± 02.96	94.66 ± 01.94
(viii) DAUNet-MS-DS-AG	0.5	78.43 ± 27.56	85.92 ± 13.73	92.99 ± 02.76	91.19 ± 04.71	95.04 ± 02.89
(ix) DAGUNet-MS-DS	0.4	81.54 ± 24.95	87.15 ± 13.34	95.24 ± 01.33	**93.52 ± 02.39**	97.06 ± 00.83
(x) DAGUNet-MS-DS-AG	0.4	80.74 ± 24.89	86.79 ± 12.00	94.78 ± 00.99	93.03 ± 01.91	96.63 ± 00.76

Bold values are used to highlight the best results within each column.

TABLE 2 | Segmentation and detection performances obtained with the four main designs considered and the nnU-Net baseline, averaged over the five-folds.

Experiment	PT	Dice	Dice-TP	F1	Recall	Precision
UNet-Slabs	0.6	74.41 ± 29.04	81.72 ± 18.19	82.74 ± 02.65	91.04 ± 03.87	75.91 ± 02.89
PLS-Net	0.5	71.69 ± 33.41	83.46 ± 17.96	89.87 ± 01.79	85.88 ± 03.02	94.31 ± 01.03
UNet-FV	0.5	76.91 ± 28.98	84.77 ± 16.22	93.19 ± 01.70	90.70 ± 01.90	95.86 ± 02.28
Ours	0.4	81.64 ± 25.33	**87.69 ± 12.12**	**95.58 ± 02.24**	93.03 ± 04.13	**98.39 ± 01.43**
nnUNet	–	**83.55 ± 21.69**	86.28 ± 7.30	86.20 ± 1.37	**96.83 ± 1.21**	77.71 ± 2.16

The first two designs were introduced and detailed in our previous study (38). Bold values are used to highlight the best results within each column.

task. The simple U-Net architecture over an entire 3D volume (UNet-FV), building upon the strengths of UNet-Slabs and PLS-Net, boosts performances in every aspect. Employing advanced mechanisms such as attention, deep supervision, or multi-scale input provides slight improvements in detection performances, going from an F1-score of 93.2 up to 95.6%. Yet, the highest benefit can be witnessed for the pixel-wise segmentation task, with an overall Dice score reaching 81.64%, up by almost 5%. Both the UNet-Slabs and nnU-Net approaches share similarities in their designs whereby subdivisions of the original MRI volumes are used. As such, both approaches obtain precision performance below 80%, far beneath the 98% from our approach. However, from its use of MRI volumes at a high resolution and internal design optimized for the dataset, the nnU-Net approach reaches the best recall performance with 96.83%. Finally, and as indicated by the Dice-TP scores, both nnU-Net and our best approach are performing similar pixel-wise segmentation for identified meningiomas.

Visual comparisons are provided in **Figure 3** between the four methods and nnU-Net baseline for six different patients, one per row. Those meningiomas were hand-picked because of their locations in uncommon regions of the brain or their small volumes. For the patients featured in the first two rows, false positive segmentations can be seen over contrast-enhanced blood vessel regions for the UNet-Slabs and nnU-Net approaches, not existing with our best approach. The patient featured in the third and fourth rows are representative for challenging meningiomas with a volume < 3 ml. In such cases, only nnU-Net and our best approach can reach reasonable Dice scores. A better pixel-wise quality can be expected from nnU-Net using a higher resolution input volume (cf. third row), which has to be weighed against its higher false positive rate compared to our AGUNet architecture (cf. fourth row). For the last two patients displayed, the meningioma s are almost completely outside the brain and have grown between the eye sockets or up the back of the nose, location relatively rare and under-represented in our dataset. From their design bringing more focus to local intensity gradients and less on overall brain location, both the UNet-slabs and nnU-Net approaches fare better. The use of more global information with the UNet-FV approach reduces prediction probabilities further away from the brain. Lastly, the use of attention mechanisms with our best approach lowers significantly the Dice score, up to a third of the nnU-Net results (cf. last row). Even though the impact of attention mechanisms could not be overall witnessed from the values reported in **Table 1**, the results on those two patients represent a perfect exemplification. While such meningiomas are found with our best approach, attention mechanisms seem to have learned to limit the predictions within the brain or its outskirts. The positive impact of attention mechanisms is then the increased specificity and better disambiguation between tumor tissue and other contrast-enhancing structures. With our AGUNet architecture almost no false positives are generated compared to nnU-Net. As illustrated in the first row of **Figure 4**, feature maps benefiting from attention are not responding on other contrast-enhancing structures, in opposition to feature maps from the UNet-FV architecture. In the second row, a higher and more selective response can be seen over the menininginoma location with our AGUNet architecture. For the UNet-FV architecture, the response is on average higher across the brain region and less specific to the meningioma location. From the training examples, as not many meningiomas in our dataset are outgrowing this far from the brain , it remains to be seen if a larger collection would improve the attention feature maps.

4.3. Volume-Based Performances Analysis

Based on tumor volume, the meningiomas from our dataset were grouped into ten equally populated bins and Dice performances for each bin are reported using a box plot, as shown in **Figure 5**. The average Dice score for the largest meningiomas, with a volume of at least 17.5 ml, has not changed much across the different methods considered and hovers above 90%. However, the number of undetected or poorly segmented large meningiomas is lessened with our best method, as can be seen by the reduced number of dots outside the whiskers of each box plot. With our best approach, we reach an overall recall of 93%, which increases to 98% considering only meningiomas larger than 3 ml. We have identified 11 undetected cases with a volume larger than 3 ml, and two examples are provided in the second row of **Figure 6**. Both a non-enhancing intraosseus meningioma (to the left) and a partly calcified meningioma (to the right) are featured. All 11 cases are exhibiting some extent of contrast impediment compared to typical contrast-enhancing meningiomas (cf. first row of **Figure 6**), which explains why our network struggles. Considering that the average meningioma in an hospital setting has a volume of 30.92 ml and the performances on meningiomas larger than 3 ml, our proposed approach appears suitable and relevant.

The most significant results of our best approach can be observed for meningiomas smaller than 3 ml, where the average

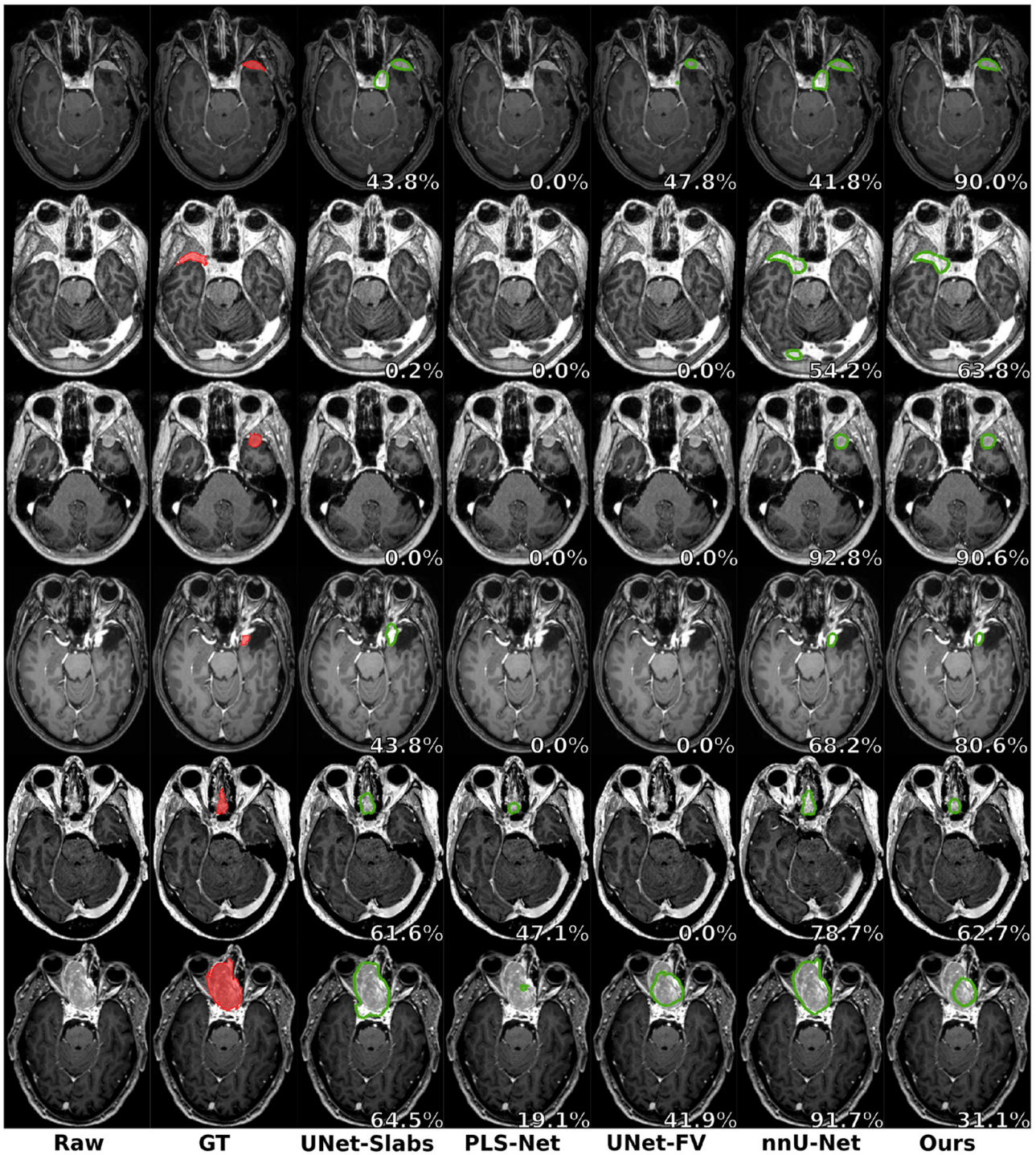

FIGURE 3 | Prediction examples for meningiomas with uncommon shapes, locations, or small volumes. One patient is featured per row and the overall raw Dice is reported in white.

Dice scores has been clearly improved. Starting from an average Dice of 46% and recall of 62.7% with PLS-Net, our best approach reaches an average Dice of 63.3% and recall of 78.9%. From its design and the use of MRI volumes at a high resolution, the nnU-Net approach has reached a Dice of 71.5% and a recall of 91% on this specific category. The difference is especially striking for meningiomas with a volume $<$ 1 ml, but for overall worse results considering larger meningiomas.

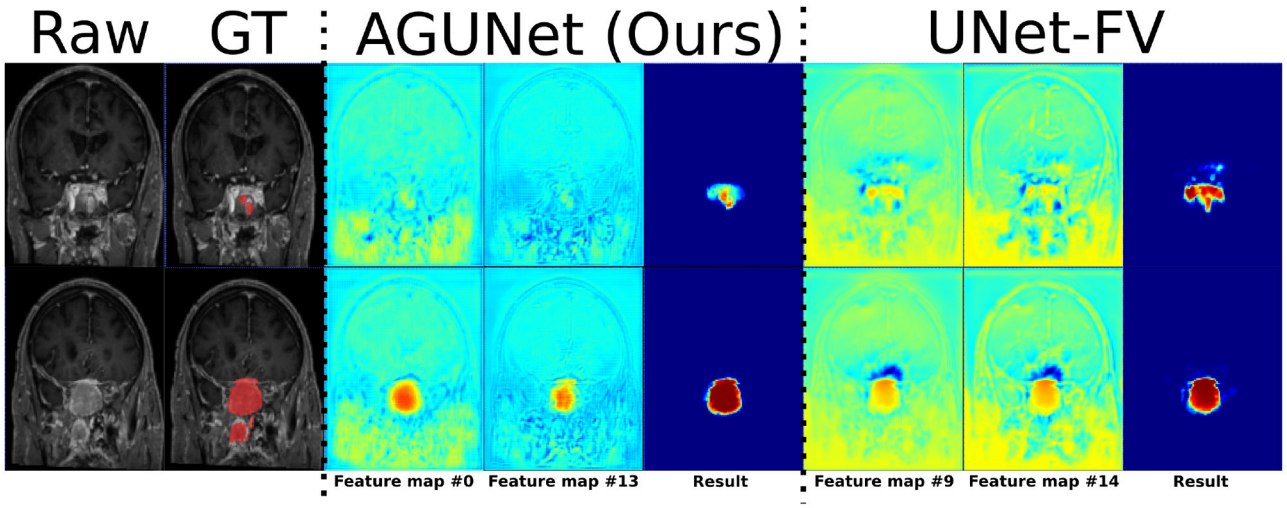

FIGURE 4 | Illustration of the effect of attention mechanisms on two feature maps from the last decoder level, between AGUNet (Ours) and UNet-FV architectures. Each row features a different slice along the coronal view of the same patient. Feature maps values are scaled in the range [0, 1], with blue to red as corresponding color code.

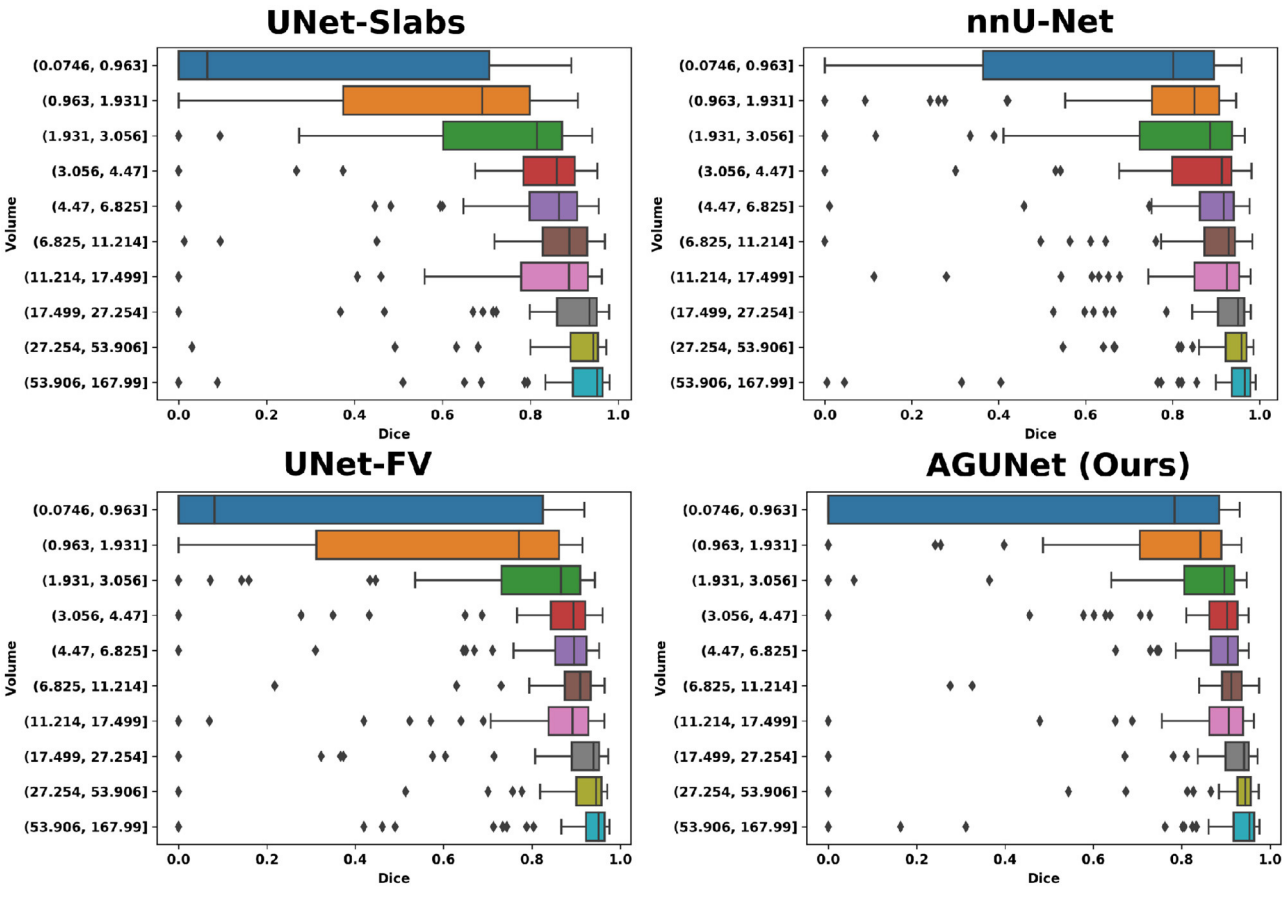

FIGURE 5 | Segmentation performances for the four main designs represented with box plots. Ten equally populated bins, based on tumor volumes, have been used to group the meningiomas.

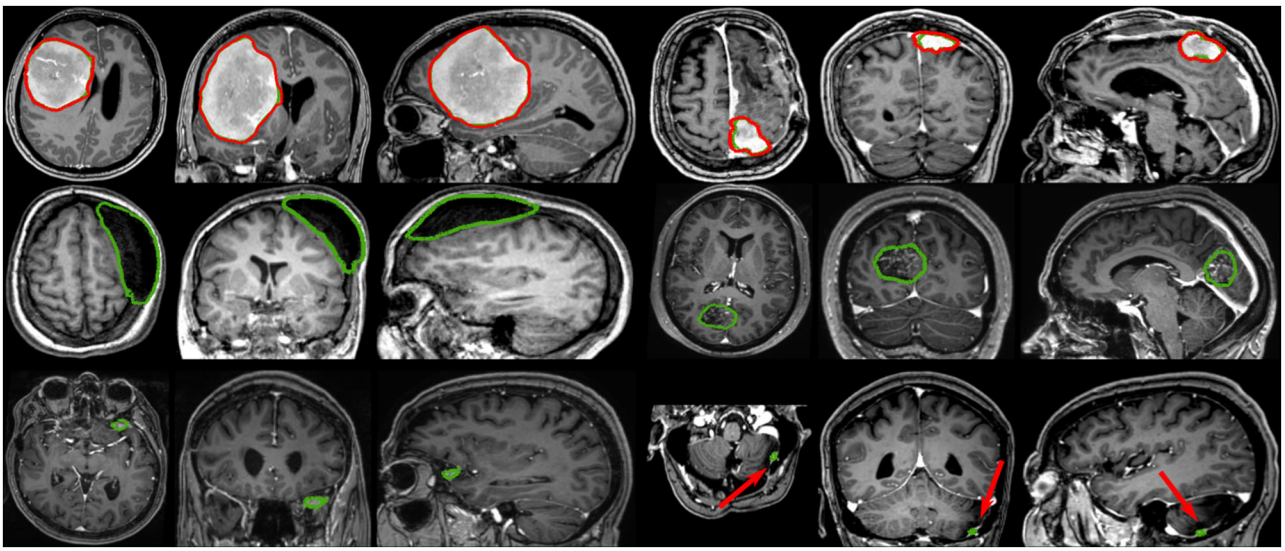

FIGURE 6 | Segmentation examples showing side-by-side the axial, coronal, and sagittal views respectively, where the automatic segmentation is shown in red and the manual annotation is shown in green. The top row illustrates two properly segmented meningiomas with to the right a meningioma adjacent to the enhancing superior sagittal sinus and falx. The middle row shows to the left a non-enhancing meningioma exhibiting intraosseus growth and hyperostosis, and to the right a partly calcified, partly enhancing meningioma. The bottom row illustrates two meningiomas, with a volume smaller than 3 ml, left undetected.

In **Figure 6**, two representative meningiomas smaller than 3 ml and left unsegmented by all methods are illustrated in the third row. Locations around brain borders (e.g., eye socket or brainstem) and close to larger blood vessels are especially challenging. In addition, using down-sampled MRI volumes with limited spatial resolution reduces such meningiomas to a very limited number of voxels the model can compute features from. Incidental findings of meningiomas' first appearance, when below 3 ml, remains challenging and unreliable for broad clinical use. However, patients followed at the outpatient clinic have developed meningiomas of 7.62 ml on average, suggesting potential benefit from automatic segmentation using our models.

4.4. Speed Performances Study

The model complexity, training convergence aspects, inference speed, and processing speed are reported in **Table 3**. Multiple GPUs with slightly different specifications (e.g., memory clock speed and bandwidth) were used for training, and other CPU-demanding algorithms were episodically ran concurrently. As a result, the speeds per epoch and total train time reported cannot be directly and objectively compared but orders of magnitude are nonetheless relevant to consider. The PLS-Net architecture is converging in less than 100 epochs, the fastest of all investigated designs. Even with the smallest number of total parameters, its complex operations result in a total training time about three times longer than any full volume U-Net design for worse segmentation performances due to its shallowness [cf. experiments (ii) and (iii)]. Training in a slab-wise fashion inherently increases the number of training samples which considerably lengthen the elapsed time per epoch by a ten-fold compared to the fastest iterating design [cf. experiments (i) and (iii)]. However, the convergence behavior is not impacted as about 120 epochs are necessary, which is on-par with the various full volume designs such as experiment (vii). It is worth noting that while using accumulated gradients does not improve overall segmentation and detection performances, the models converge faster thanks to a better generalization from seeing more samples at every epoch [cf. experiments (vi) and (vii)]. The combination of complex architectural designs and accumulated gradients enables convergence in about 110 epochs at best, which is equivalent to a more than reasonable total training time of 18 h. One must trade carefully between model complexity and dataset size to prevent overfitting or similar convergence hurdles. The use of full volume inputs, the relatively small dataset size, and the quickly increasing total number of model parameters with advanced designs are complex to balance.

Regarding inference, doubling the number of parameters within a similar architecture does not alter the speed as can be seen between experiments (iv) and (ix). Yet, only the shallow architecture from PLS-Net can go below the second. When running experiment (ix) on CPU, the inference speed reaches on average 8.66 ± 0.09 s, slightly more than doubled compared to GPU usage. The largest gap between CPU and GPU usage happens when running experiment (ii). With regard to the total processing time for a new patient's MRI volume, around 15 s are necessary to provide segmentation predictions using a GPU, which would be fast enough not to hinder day-to-day clinical practice. Interestingly, and when considering computers deprived of high-end GPUs, the processing time on CPU remains similar with 15.39 ± 0.15 s for experiment (ix). When running inference on GPU for only one patient, the environment has to be initialized at first and the model loaded, making it speed-wise

TABLE 3 | Model complexity, training convergence, and runtime performances for the different architecture designs studied, averaged across the five-folds.

Experiment	No. of params (M)	s.*epoch*$^{-1}$ (s)	Best epoch	Train time (h)	Inference (s)	Processing (s)
(i) UNet-Slabs (38)	14.75	$4,103 \pm 313$	120 ± 40	160.2 ± 44.3	3.74 ± 0.03	15.53 ± 0.16
(ii) PLS-Net (38)	0.25	$1,944 \pm 47$	91 ± 23	62.6 ± 12.5	0.92 ± 0.01	10.95 ± 0.05
(iii) UNet-FV	5.89	374 ± 4.7	171 ± 24	21.0 ± 02.5	2.03 ± 0.04	12.12 ± 0.08
(iv) AGUNet-AG	16.41	437 ± 3.6	138 ± 28	20.5 ± 03.5	3.88 ± 0.04	13.84 ± 0.13
(v) AGUNet-DS-AG	16.41	434 ± 3.4	149 ± 27	21.7 ± 03.3	3.59 ± 0.06	14.13 ± 0.21
(vi) AGUNet-MS-DS	18.66	472 ± 3.5	160 ± 78	25.0 ± 10.3	3.69 ± 0.04	14.35 ± 0.15
(vii) AGUNet-MS-DS-AG	18.66	508 ± 8.1	120 ± 29	21.4 ± 04.1	3.71 ± 0.04	14.46 ± 0.22
(viii) DAUNet-MS-DS-AG	25.72	434 ± 3.3	118 ± 52	18.0 ± 06.2	3.13 ± 0.04	13.85 ± 0.17
(ix) DAGUNet-MS-DS-AG	30.96	476 ± 3.3	112 ± 14	18.9 ± 01.9	3.32 ± 0.06	16.13 ± 0.28

All values in the table are reported with GPU support. Models trained with the nnU-Net architecture are made of 25.25 M parameters.

comparable with pure CPU usage. The serious bottlenecks when using computers with average specifications could be the RAM availability and the CPU parameters (e.g., frequency or number of cores).

5. DISCUSSION

In this study, we investigated different deep learning architectures and designs for segmenting meningiomas in T1-weighted MRI volumes, relying on attention mechanisms and global relationships. Slab-wise and patch-wise approaches, poorly benefiting from global information, ostensibly struggle to reach high F1 scores due to the generation of many false positives. Locally, all hyperintense structures appear quite similar to one another. Directly leveraging an entire 3D volume, even with simple architectures such as U-Net, has the clear benefit of dramatically reducing the number of false positives per patient. Having access to global context and spatial relationships across the whole brain helps the model to better discriminate between contrast-enhanced meningiomas and bright anatomical structures (e.g., blood vessels). Interestingly, an improved modeling of spatial relationships has close to no positive effect on the pixel-wise segmentation quality. Actually, the lack of satisfactory spatial resolution from the use of a down-sampled input volume can prove to be detrimental. In order to boost recall performance as high as possible, especially on small meningiomas, an efficient use of high-resolution network inputs is necessary as proven by the nnU-Net approach. A joint generation of better global context features, preservation of local information, and leveraging of MRI volumes at their native resolution is key to push performances higher. Alternatively, performing some extent of ensembling could bear potential. The strengths from each approach could be built upon while inhibiting limitations regarding precision and pixel-wise segmentation accuracy. However, improved segmentation performance would come at the expense of speed performance and additional complexity. In addition, increasing the amount of models in the ensemble will linearly increase the training and inference computation time.

By extending a regular U-Net backbone architecture with various designs, we managed to further improve segmentation and detection performances. However, the only noticeable and clear contribution seems to come from the use of deep supervision. Setting trainable weights in the loss function to let the model learn how to best balance the loss from the probability map at each decoder level has not been attempted in this study. We hypothesize overweighting the coarse feature maps might favor recall while overweighting the fine feature maps would favor pixel-wise segmentation, and believe further investigation is of interest. Not supported by numbers, the effect of attention schemes has been qualitatively observed whereby predictions appear to be restricted to the brain itself or its close boundaries. From training examples, the model learned global spatial relationships to define some no-prediction zones where meningiomas are unlikely to occur. While such observations warrant a proper behavior from the use of attention schemes, a greater variability in the training samples to feature meningiomas in all possible location might also be implied. Conversely, having witnessed some extent of brain-clipping effect using attention mechanisms can be considered as an indication for unsuitability toward meningioma segmentation. Given the possibility for meningiomas to potentially grow outward from every border of the brain, heavier preprocessing such as brain-masking used for glioma segmentation is inadvisable here as it would clip away parts of the tumor. The use of multi-scale inputs also brought limited visible improvement, but the training samples fed to our architectures were already down-sampled from the original 3D MR volumes, starting the training with a degraded spatial resolution. For the time being, training our best architecture with the native MRI volume resolution is too challenging because of memory limitation on even high-end GPUs due to the sizable memory footprint. Nonetheless, working with down-sampled input volumes seems like the best trade-off solution as both recall and precision are favored. Detecting each and every meningioma accurately is critical as the actual pixel-wise segmentation task is more than often eased by the relatively good contrast and non-diffuse aspect of such tumors. Overall, the total amount of trainable parameters for a model should be considered when assessing performances. Architectures with deeper or wider designs are more prone to outperforming lightweight architectures.

To this day, the dataset featured in this study is the largest used for the task of meningioma segmentation and includes a wider range of tumor characteristics (e.g., volume

and location). The joint effect of the proposed architectures coupled to the diverse dataset improves over the state-of-the-art results for meningiomas. The current segmentation and detection performances are exhibiting a satisfactory potential for clinical use either as a tool for surgical planning or growth estimation. Automatic measurements regarding the tumor aspect (i.e., volume, short-axis) and location (i.e., brain hemisphere and lobe) can be automatically generated. An average Dice score above 90% was reached for meningiomas bigger than 3 ml, when the average volume for patients having undergone surgery is 30.92 ml. For meningiomas with a volume below 3 ml, somewhat worse performances were obtained. Detection of early meningiomas appears to be feasible but further improvements are needed for real and trustworthy use. As the average volume from patients followed at the outpatient clinic is 7.62 ml, the current performances open for automatic and systematic growth computation during follow-up over time. In addition, inter-/intra-observer variability would be lessened, as well as time consumption for clinicians. By providing an open access to our trained models and inference scripts, other research groups should have an easier time to put together their own annotated datasets. In turn, new studies on the task of pixel-wise meningioma segmentation or more clinically oriented could be fostered.

The utmost challenging task remains the detection of tiny meningiomas exhibiting visual similarities with blood vessels, sometimes placed side-by-side or overlapping with them. The smallest meningiomas are also featured in a wider range of location (e.g., along the brainstem), and their total volume is only represented by a handful of voxels given the initial volume down-sampling. To address shortcomings from the latter, a finer down-sampling would help retain a superior spatial resolution but finding the proper balance between memory requirement and a prominent risk of overfitting would be challenging. Furthermore, broadening the dataset with additional samples featuring small meningiomas in a vaster range of locations might help the trained models generalize better. Alternatively, the use of other MR sequences such as FLAIR could help better distinguish between tumor and vessels. However, a larger panel of MR sequences might not be available at all time and processing only T1-weighted volumes makes our approach more generic and easier to use. Lastly, improving the architecture to make a better use of features available at the different input scales might be considered.

To allow for exact comparison with the results from our previous study (38), the dataset was not altered after the identification of outliers where meningiomas would not show with proper contrast, and which could be considered to be excluded from future studies. Discarding the 98 T1-weighted MRI volumes with a slice thickness > 2.0 mm from the original dataset was a study choice. For diagnosis, only 3D MR scans with a slice thickness up to 2 mm are holding relevant information for visual inspection. In a previous study (38), similar performances were obtained whether the low-resolution MRI volumes were included in the training set or not, not impeding the ability of a network to be efficiently trained. Finally, forcing an intense image-stretching during preprocessing to reach the 1.0 mm^3 spacing leads to heavy blurring, which is detrimental for features computation and interferes in the details preservation and feature maps quality improvement of the architectures. Different or adaptive preprocessing approaches would need to be further investigated. As it stands, the 11 outliers out of 600 volumes are additional noise during training and are a hindrance for the training process. By excluding them during validation, we would virtually reach 100% recall with our best-performing model for meningiomas bigger than 3 ml. In the validation studies, we chose to only rely on the threshold value PT, applied over the prediction map, to report the segmentation performance results. With the different full volume approaches, an almost perfect precision and high Dice scores were obtained. As a consequence, using an additional detection threshold was not deemed necessary, whereby a true positive is acknowledged only if the Dice score is above the given threshold. Only few meningiomas have poor pixel segmentation and extent coverage (i.e., Dice score below 50%), while the near-perfect precision ascertains the detection to be at least part of a meningioma.

Even with sophisticated architectures and heavier designs, models are extremely fast to train and are converging in under 20 h. Using an entire 3D volume as input compared to a slabbing strategy also speeds up training as less training samples are processed during each epoch. In addition, generalization schemes such as accumulated gradients help the model converge faster and reach a better optimum as can be seen by the reduction in standard deviation for the segmentation and detection measurements. Interestingly, the inference speed is not heavily impacted by large variations in model complexity and these two parameters do not linearly correlate. Our dual attention guided architecture has 100 times more parameters than the shallow PLS-Net architecture. Yet, the inference speed is only multiplied by 3 reaching at most 3.7 s which is still fast enough and relevant for clinical use. The biggest hurdle for deployment in hospitals would be the large variability in hardware from low/mid-end computers and where shallower architectures like PLS-Net could thrive. The current disparity in performances, around 6% F1-score difference, remains too high for such consideration at the moment and further investigation in that direction is warranted.

In the future, the focus should be on improving the segmentation performance for meningiomas with a volume under 3 ml. A new round of data collection should be performed, especially from the outpatient clinic were such meningiomas are more heavily represented. Conjointly, experiments using multi-scale concepts introduced in section 2.1 should be carried out. By computing features across a wider range of scales, more global knowledge could be gained, without deteriorating the pixel-wise segmentation quality. Blood vessels represent the main source of confusion for the models, since looking very similar to the smallest meningiomas. A separate way to handle blood vessels might then be of interest to perform disambiguation from tumors, or to refine the segmentation. However, no dataset exists for the task and the time required to manually segment blood vessels is prohibitive. Finally, the use of advanced loss functions to refine the segmentation around the tumors' edges is appealing. The Dice coefficient as loss function is unable to capture all tumor's aspects and favors large main tumors at the detriment of off-sites. A combined loss function including surface-distance metrics or instance-wise metrics should be further investigated.

6. CONCLUSION

In this paper, we pushed forward the investigations around spatial relationships and global context for the task of meningioma segmentation in T1-weighted MRI volumes. Integrated into a regular U-Net backbone, we experimented with concepts such as attention mechanisms, multi-scale input, and deep supervision. Improved segmentation and detection performances have been demonstrated when moving from slab-wise to more sophisticated and complex approaches leveraging the entire 3D volume. Almost perfect detection results for clinically relevant meningiomas were obtained. On the other hand, the smallest meningiomas, with a volume below 3 ml, remained challenging given the limited spatial resolution and limited number of voxels to compute features from. In future work, special care should be brought toward the training dataset, as in many applications the bottleneck for improving performances lies in the data diversity more than the method's design (56). Nevertheless, smarter handling of multi-scale features should be investigated, such as spatial pyramid pooling, to better leverage the raw spatial resolution. Alternative loss function designs, using adaptive weighting or new concepts, might also improve the pixel-wise segmentation, especially around tumor borders.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: patient data are protected under GDPR and cannot be distributed. Requests to access these datasets should be directed to David Bouget, david.bouget@sintef.no.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

DB and AP contributed to conception and design of the study. SH and OS contributed to the data acquisition and labeling process. DB, AP, IR, and OS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was funded by the Norwegian National Advisory Unit for Ultrasound and Image-Guided Therapy (usigt.org).

REFERENCES

1. Bauer S, Wiest R, Nolte LP, Reyes M. A survey of MRI-based medical image analysis for brain tumor studies. *Phys Med Biol*. (2013) 58:R97. doi: 10.1088/0031-9155/58/13/R97
2. Ostrom QT, Cioffi G, Gittleman H, Patil N, Waite K, Kruchko C, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012-2016. *Neuro-oncology*. (2019) 21(Suppl. 5):v1–100. doi: 10.1093/neuonc/noz150
3. Vernooij MW, Ikram MA, Tanghe HL, Vincent AJ, Hofman A, Krestin GP, et al. Incidental findings on brain MRI in the general population. *N Engl J Med*. (2007) 357:1821–8. doi: 10.1056/NEJMoa070972
4. Solheim O, Torsteinsen M, Johannesen TB, Jakola AS. Effects of cerebral magnetic resonance imaging in outpatients on observed incidence of intracranial tumors and patient survival: a national observational study. *J Neurosurg*. (2014) 120:827–32. doi: 10.3171/2013.12.JNS131312
5. Goldbrunner R, Minniti G, Preusser M, Jenkinson MD, Sallabanda K, Houdart E, et al. EANO guidelines for the diagnosis and treatment of meningiomas. *Lancet Oncol*. (2016) 17:e383–91. doi: 10.1016/S1470-2045(16)30321-7
6. Berntsen EM, Stensjøen AL, Langlo MS, Simonsen SQ, Christensen P, Moholdt VA, et al. Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. *Acta Neurochirur*. (2020) 162:379–87. doi: 10.1007/s00701-019-04110-0
7. Binaghi E, Pedoia V, Balbi S. Collection and fuzzy estimation of truth labels in glial tumour segmentation studies. *Comput Methods Biomech Biomed Eng*. (2016) 4:214–28. doi: 10.1080/21681163.2014.947006
8. Behbahani M, Skeie GO, Eide GE, Hausken A, Lund-Johansen M, Skeie BS. A prospective study of the natural history of incidental meningioma-Hold your horses! *Neuro-Oncol Pract*. (2019) 6:438–50. doi: 10.1093/nop/npz011
9. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imag*. (2010) 29:1310–20. doi: 10.1109/TMI.2010.2046908
10. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imag*. (2000) 19:143–50. doi: 10.1109/42.836373
11. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. (2017) 39:2481–95. doi: 10.1109/TPAMI.2016.2644615
12. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2015). p. 3431–40. doi: 10.1109/CVPR.2015.7298965
13. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28
14. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Human-level CMR image analysis with deep fully convolutional networks (2017). *arXiv preprint arXiv:1710.09289v1*.
15. Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. *IEEE Trans Neural Netw Learn Syst*. (2019) 30:3484–95. doi: 10.1109/TNNLS.2019.2892409
16. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. *J Digit Imag*. (2017) 30:449–59. doi: 10.1007/s10278-017-9983-4
17. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:170605587*. (2017). doi: 10.1007/978-3-030-01234-2_49
18. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 2881–90. doi: 10.1109/CVPR.2017.660
19. Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters-improve semantic segmentation by global convolutional network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 4353–61. doi: 10.1109/CVPR.2017.189
20. Lin G, Milan A, Shen C, Reid I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 1925–34. doi: 10.1109/CVPR.2017.549

21. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2019). p. 3146–54. doi: 10.1109/CVPR.2019.00326
22. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:180403999*. (2018).
23. Cheng J, Tian S, Yu L, Lu H, Lv X. Fully convolutional attention network for biomedical image segmentation. *Artif Intell Med*. (2020) 2020:101899. doi: 10.1016/j.artmed.2020.101899
24. Feng X, Tustison NJ, Patel SH, Meyer CH. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Front Comput Neurosci*. (2020) 14:25. doi: 10.3389/fncom.2020.00025
25. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, et al. Bach: Grand challenge on breast cancer histology images. *Med Image Anal*. (2019) 56:122–39. doi: 10.1016/j.media.2019.05.010
26. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imag*. (2014) 34:1993–2024. doi: 10.1109/TMI.2014.2377694
27. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. (2017) 35:18–31. doi: 10.1016/j.media.2016.05.004
28. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med Image Anal*. (2018) 43:98–111. doi: 10.1016/j.media.2017.10.002
29. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. (2016) 35:1240–51. doi: 10.1109/TMI.2016.2538465
30. Dvorak P, Menze B. Structured prediction with convolutional neural networks for multimodal brain tumor segmentation. In: *Proceeding of the Multimodal Brain Tumor Image Segmentation Challenge*. (2015) p. 13–24. doi: 10.1007/978-3-319-42016-5_6
31. Zikic D, Ioannou Y, Brown M, Criminisi A. Segmentation of brain tumor tissues with convolutional neural networks. In: *Proceedings MICCAI-BRATS*. (2014). p. 36–9.
32. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: *International MICCAI Brainlesion Workshop*. Springer (2018). p. 311–20. doi: 10.1007/978-3-030-11726-9_28
33. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. No new-net. In: *International MICCAI Brainlesion Workshop*. Springer (2018). p. 234–44. doi: 10.1007/978-3-030-11726-9_21
34. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. (2017) 36:61–78. doi: 10.1016/j.media.2016.10.004
35. Xu Y, Gong M, Fu H, Tao D, Zhang K, Batmanghelich K. Multi-scale masked 3-D U-net for brain tumor segmentation. In: *International MICCAI Brainlesion Workshop*. Springer (2018). p. 222–33. doi: 10.1007/978-3-030-11726-9_20
36. Laukamp KR, Thiele F, Shakirin G, Zopfs D, Faymonville A, Timmer M, et al. Fully automated detection and segmentation of meningiomas using deep learning on routine multiparametric MRI. *Eur Radiol*. (2019) 29:124–32. doi: 10.1007/s00330-018-5595-8
37. Laukamp KR, Pennig L, Thiele F, Reimer R, Görtz L, Shakirin G, et al. Automated meningioma segmentation in multiparametric MRI. *Clin Neuroradiol*. (2020) 31:357–66. doi: 10.1007/s00062-020-00884-4
38. Bouget D, Pedersen A, Hosainey SAM, Vanel J, Solheim O, Reinertsen I. Fast meningioma segmentation in T1-weighted magnetic resonance imaging volumes using a lightweight 3D deep learning architecture. *J Med Imaging*. (2021) 8:024002. doi: 10.1117/1.JMI.8.2.024002
39. Sang H, Zhou Q, Zhao Y. PCANet: Pyramid convolutional attention network for semantic segmentation. *Image Vis Comput*. (2020) 103:103997. doi: 10.1016/j.imavis.2020.103997
40. Jetley S, Lord NA, Lee N, Torr PHS. Learn to pay attention. *arXiv:1804.02391*. (2018).
41. Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, et al. A structured self-attentivesentence embedding. *arXiv preprint arXiv:170303130*. (2017).
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. (2017). p. 5998–6008.
43. Lin G, Shen C, Van Den Hengel A, Reid I. Efficient piecewise training of deep structured models for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016). p. 3194–203. doi: 10.1109/CVPR.2016.348
44. Abraham N, Khan NM. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. (2019). p. 683–7. doi: 10.1109/ISBI.2019.8759329
45. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, et al. Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal*. (2019) 53:197–207. doi: 10.1016/j.media.2019.01.012
46. Mou L, Zhao Y, Chen L, Cheng J, Gu Z, Hao H, et al. CS-Net: channel and spatial attention network for curvilinear structure segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* Springer (2019). p. 721–30. doi: 10.1007/978-3-030-32239-7_80
47. Sinha A, Dolz J. Multi-scale self-guided attention for medical image segmentation. *IEEE J Biomed Health Informatics*. (2020) 25:121–30. doi: 10.1109/JBHI.2020.2986926
48. Niu R. Hmanet: Hybrid multiple attention network for semantic segmentation in aerial images. *arXiv preprint arXiv:200102870*. (2020).
49. Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: *Artificial Intelligence and Statistics*. (2015). p. 562–70.
50. Wang F, Eljarrat A, Müller J, Henninen TR, Erni R, Koch CT. Multi-resolution convolutional neural networks for inverse problems. *Sci Rep*. (2020) 10:1–11. doi: 10.1038/s41598-020-62484-z
51. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2018). p. 7794–803. doi: 10.1109/CVPR.2018.00813
52. Masters D, Luschi C. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:180407612*. (2018).
53. Killeen PR. An alternative to null-hypothesis significance tests. *Psychol Sci*. (2005) 16:345–53. doi: 10.1111/j.0956-7976.2005.01538.x
54. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:180910486*. (2018). doi: 10.1007/978-3-658-25326-4_7
55. Jung AB, Wada K, Crall J, Tanaka S, Graving J, Reinders C, et al. *Imgaug*. (2020). Available online at: https://github.com/aleju/imgaug (accessed February 1, 2020.)
56. Hofmanninger J, Prayer F, Pan J, Rohrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem. *arXiv preprint arXiv:200111767*. (2020). doi: 10.1186/s41747-020-00173-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.