Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Anatomically-Guided Attention for False Positive Reduction in Brain Tumor Segmentation

Kashvi Ruparelia

kashvi1729@gmail.com

Anand Niketan Satellite https://orcid.org/0009-0006-7280-0367

Jayendra Bhalodiya

Ahmedabad University

Research Article

Keywords: Brain tumor segmentation, U-Net, attention, anatomical prior, false positives, BRISC

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11603150/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Anatomically-Guided Attention for False Positive Reduction in Brain Tumor Segmentation

Kashvi Ruparelia
Anand Niketan Satellite
Ahmedabad, India

Dr. Jayendra Bhalodiya
Assistant Professor, School of Engineering and Applied Science
Ahmedabad University
Ahmedabad, India

***Abstract*—U-Net-based brain tumor segmentation can produce false positives in anatomically implausible locations, which inflate tumor volume estimates and reduce clinical trust. We investigate whether injecting an anatomical location prior into the network's attention can reduce these errors. We propose a prior-informed attention gate that conditions U-Net features on empirical tumor-location prior maps computed from the training annotations, applied as a residual spatial gate at the bottleneck and at every decoder stage. Using the BRISC dataset of 2D contrast-enhanced T1-weighted MRI slices (glioma, meningioma and pituitary tumor), we compare the proposed model with a vanilla U-Net and a learned-gate ablation under identical training, using 3-fold cross-validation and a held-out test set of 860 images. We evaluate Dice, IoU, precision, recall, Hausdorff distance, HD95 and a dedicated far false-positive measure. The proposed model matched the baseline in Dice (0.709 versus 0.708) and showed a slightly lower but non-significant HD95 (36.5 versus 37.9 px, p = 0.26); it did not significantly reduce far false positives (49.7 versus 48.0 px, p = 0.82). A simple location prior injected through a spatial gate is therefore insufficient on BRISC, and stronger prior integration, longer training and additional datasets are needed to establish any benefit.**

***Index Terms*—*Brain tumor segmentation, U-Net, attention, anatomical prior, false positives, BRISC.*

I. INTRODUCTION

Accurate segmentation of brain tumors from MRI is a prerequisite for reliable diagnosis, treatment planning, and monitoring in neuro-oncology. Deep learning-based segmentation, particularly U-Net and its variants, has become the dominant approach to solve this problem [1].

Progress in the field has historically been driven by benchmark datasets such as BraTS [2], which established multimodal, expert-annotated glioma segmentation as a standard task. However, BraTS-style datasets are limited in scope: they focus almost exclusively on gliomas, depend on heavily pre-processed and harmonized volumes that may not reflect real-world acquisition variability, and provide limited diversity in tumor type [3]. The recently introduced BRISC dataset [4] addresses several of these gaps. BRISC comprises 6,000 contrast-enhanced T1-weighted MRI images spanning four diagnostic categories — glioma, meningioma, pituitary tumor, and non-tumorous cases — captured across axial, sagittal, and coronal imaging planes, with 4,793 images carrying pixel-wise expert-verified segmentation masks. Because BRISC covers multiple tumor types rather than gliomas alone, and includes a non-tumorous class, it offers a more demanding and clinically representative setting in which to study a specific and consequential failure mode of segmentation models: false positive predictions in anatomically implausible locations.

Convolutional architectures such as U-Net learn spatial patterns primarily from local intensity and texture cues, which makes them prone to misclassifying tumor-mimicking structures — vascular tissue, imaging artifacts, or normal anatomical boundaries — as tumor, particularly when such regions do not conform to the typical spatial distribution of tumors within the brain. This is problematic clinically, since false positives can inflate estimated tumor volume and reduce clinician trust in automated tools. Attention mechanisms improve long-range spatial context modeling, but standard attention does not explicitly encode anatomical priors about where tumors are and are not likely to occur, and so remains susceptible to similar false positive patterns.

This project investigates whether incorporating anatomical priors directly into the attention mechanism of a U-Net-based segmentation model can reduce false positive predictions on the BRISC dataset without sacrificing sensitivity to true

tumor regions. We design a prior-informed attention module that biases the network's attention maps using tumor-location priors derived empirically from BRISC's training annotations, and evaluate it against the vanilla U-Net. Our contributions are threefold: (1) a systematic baseline evaluation of U-Net segmentation performance on BRISC using Dice Score, Intersection over Union (IoU), and Hausdorff distance; (2) the design and integration of a prior-informed attention module into the U-Net to suppress anatomically implausible activations; and (3) a comparative and ablation study isolating the contribution of the proposed module relative to the vanilla U-Net and a learned-gate variant.

II. METHODOLOGY

A. Dataset

We use the BRISC dataset [4], which consists of 2D, contrast-enhanced T1-weighted MRI slices collected and curated from multiple public sources, with expert-verified pixel-wise segmentation masks for glioma, meningioma, and pituitary tumors. Images are distributed across axial, sagittal, and coronal planes, with a segmentation subset of 4,793 paired image–mask files. We use the dataset's provided train/test split (3,933 training and 860 test images); the training split is further partitioned into folds for cross-validation (Section V). Preprocessing consists of resizing every slice to 224×224 pixels (bilinear interpolation for images, nearest-neighbor for masks) and normalizing intensities with the mean and standard deviation of the training set.

Note: BRISC provides single-sequence (T1-weighted) 2D MRI slices rather than co-registered multi-modal 3D volumes. Accordingly, this project explicitly focuses on 2D, single-modality segmentation, and the investigation will be scoped to 2D methods rather than 3D U-Net architectures or multi-modal fusion.

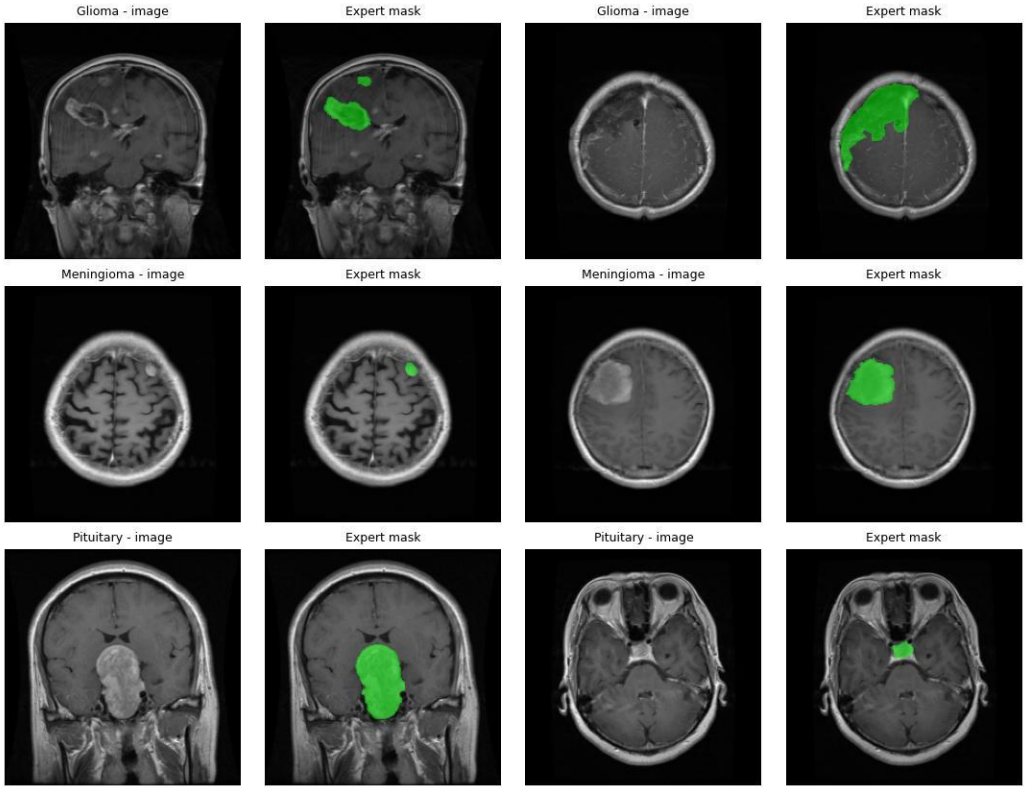


Fig. 1. Example BRISC test slices of glioma, meningioma and pituitary tumors (two per class). Each MRI slice is shown beside the same slice with the expert segmentation mask overlaid in green.

B. Problem Formulation

Given an input MRI slice $x \in \mathbb{R}^{H\times W\times 1}$ from the BRISC dataset, the objective is to learn a mapping $f_\theta : x \rightarrow \hat{y}$, where $\hat{y} \in [0,1]^{H\times W\times K}$ is a per-pixel class-probability map over K classes which are: background, glioma, meningioma, and pituitary tumor. The predicted segmentation mask is obtained as $\hat{Y}(i,j) = \text{argmax}_k\ \hat{y}_k(i,j)$ for each pixel location (i,j). This reframes

segmentation as dense, per-pixel classification rather than single-label image classification, i.e. a standard classifier learns Image to label, whereas segmentation learns Image to label per pixel.

C. Baseline U-Net Architecture

The baseline follows the standard 2D U-Net encoder–decoder structure with skip connections [1]. At the finest granularity, every unit in the network is a neuron computing a weighted sum followed by a nonlinearity,

$$z = \sum_i w_i x_i + b, \qquad a = f(z) \tag{1}$$

In a convolutional layer this weighted sum is applied locally and with shared weights: a k×k kernel slides over the input feature map, and at each spatial location computes

$$(F * K)(i,j) = \sum_{u,v} F(i+u,\, j+v)\, K(u,v) \tag{2}$$

For an input of size H×W, kernel size k, padding p, and stride s, the output spatial size is

$$H_{\text{out}} = \left\lfloor \frac{H + 2p - k}{s} \right\rfloor + 1 \tag{3}$$

(and analogously for W). We use k=3, p=1, s=1 inside each block (preserving resolution) and k=2, s=2 for down-sampling, so that Eq. (3) gives $H_{out} = H/2$ at each pooling stage. Each convolutional block has two convolution layers as mentioned in Eq. (4).

$$\hat{F} = \mathrm{LReLU}\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}(F))\big), \quad F' = \mathrm{LReLU}\big(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\hat{F}))\big) \tag{4}$$

where BN(·) is batch normalization and LReLU is Leaky ReLU,

$$\mathrm{LReLU}(x) = \max(\alpha x,\, x), \qquad \alpha = 0.1 \tag{5}$$

which avoids the "dying ReLU" problem of standard ReLU by allowing a small non-zero gradient when x<0. Following each block, 2×2 max-pooling with stride 2 is applied,

$$P(i,j) = \max_{(u,v)\,\in\,\text{window}(i,j)} F(u,v) \tag{6}$$

which halves the spatial resolution (Eq. 3), reduces computation, and introduces a degree of local spatial invariance. Channel depth is doubled at each stage to compensate. Given input x, the encoder produces a sequence of feature maps $F_1, F_2, \ldots, F_L$ across L downsampling stages, terminating in a bottleneck representation F_L with spatial size $H/2^L \times W/2^L$ and C_L channels. Dropout is applied after selected blocks during training to reduce overfitting.

The decoder mirrors the encoder: at each stage, features are upsampled via transposed convolution,

$$U_l = \mathrm{ConvT}_{2\times2}(F_{l+1}), \qquad D_l = \mathrm{Block}\big([\,U_l\,;\,S_l\,]\big) \tag{7}$$

where S_l is the skip-connection feature from the corresponding encoder stage, $[\cdot;\cdot]$ denotes channel-wise concatenation, and Block(·) applies the same convolution–normalization–activation pair as Eq. (4). The skip connection exists because downsampling in the encoder inevitably discards fine spatial detail; concatenating the encoder's pre-pooling features lets the decoder recover pixel-accurate boundaries that could not be reconstructed from the bottleneck representation alone — for segmentation, knowing what a region is does not suffice without knowing precisely where its boundary lies. A final 1×1 convolution maps the last decoder stage to K output channels, followed by a softmax to produce ŷ.

D. Prior-Informed Attention Module

To reduce false positives in anatomically implausible locations, we compute an empirical spatial prior over the training set and use it to gate the network's intermediate features. For each non-background class c, the prior map is estimated as

$$P_c(i,j) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[Y^{(n)}(i,j) = c\right] \tag{8}$$

where $Y^{(n)}$ is the ground-truth mask of the n-th training image (all images resized to a common H×W grid) and N is the number of training samples. This yields a fixed prior tensor $P \in [0,1]^{H\times W\times(K-1)}$, computed once before training and held constant.

At a chosen decoder stage with feature map $F \in \mathbb{R}^{h\times w\times C}$, the prior is resized to (h,w) via bilinear interpolation, $\tilde{P}$ = Resize(P), and used to compute a spatial attention gate:

$$A = \sigma\big(\mathrm{Conv}_{1\times1}([F\,;\,\tilde{P}])\big) \in [0,1]^{h\times w\times 1} \tag{9}$$

$$F' = F \odot A + F \tag{10}$$

where σ(·) is the sigmoid function and ⊙ denotes element-wise (broadcast) multiplication. Equation (10) is a residual gate: the network can learn to suppress activations in low-prior regions (small A) while a direct skip term prevents the prior from fully blocking gradient flow through regions the prior under-represents. The module is inserted at the bottleneck and at each decoder stage, so that both coarse semantic features and fine-grained spatial detail are conditioned on anatomical plausibility.

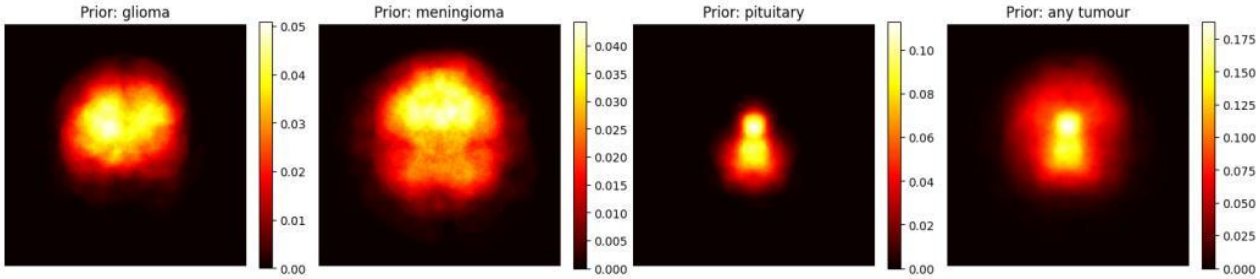


Fig. 2. Empirical anatomical prior maps (Eq. 8) computed from the training annotations: per-pixel frequency of glioma, meningioma and pituitary tumors, and of any tumor. Brighter regions mark locations where tumors occur more often; these maps are the spatial prior supplied to the prior-informed attention gate.

E. Comparison Models

The proposed prior-informed U-Net is compared against the vanilla U-Net, i.e., the baseline architecture of Section II-C without any attention mechanism. The vanilla U-Net uses the same encoder–decoder structure, convolutional blocks (Eq. 4), skip connections (Eq. 7), loss function, optimizer, data augmentation and cross-validation folds as the proposed model, so the only difference between the two is the prior-informed gate (Eqs. 9–10) at the bottleneck and at each decoder stage. This controlled design attributes any difference in segmentation accuracy or false-positive behavior to the gate itself rather than to differences in model capacity or training recipe: each gate is a single 1×1 convolution, so the gates add fewer than 500 parameters to the 1.93 M-parameter network.

As a second reference, the learned-gate U-Net uses the same gate formulation but without the anatomical prior, $A = \sigma(\mathrm{Conv}_{1\times1}(F))$, so that its attention depends only on the features themselves. Comparing the proposed model with both references separates the effect of the prior from the effect of attention-style gating in general (Section VI). All three models are trained and evaluated under identical conditions, which makes the vanilla U-Net the natural yardstick for deciding whether the anatomical prior provides any benefit.

III. LOSS FUNCTION

Training follows the standard supervised loop: each input is paired with a ground-truth mask, the network produces a prediction, the error between prediction and ground truth is scored by a loss function, and that error is backpropagated to update the weights — input, prediction, comparison with ground truth, loss computation, backpropagation, weight update, repeat.

The loss combines soft Dice loss and pixel-wise cross-entropy, following [5], to handle class imbalance and the small size of many tumor regions. It is formulated as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2}{K}\sum_{k=1}^{K}\frac{\sum_i G_{i,k}\hat{y}_{i,k}}{\sum_i G_{i,k}^2 + \sum_i \hat{y}_{i,k}^2} \tag{11}$$

where K denotes the number of classes, G and ŷ are the one-hot ground-truth label and predicted probability respectively, and i indexes pixels. This is paired with pixel-wise cross-entropy,

$$\mathcal{L}_{\text{CE}} = -\frac{1}{HW}\sum_{i=1}^{HW}\sum_{k=1}^{K} G_{i,k}\log(\hat{y}_{i,k}) \tag{12}$$

and the two are combined as

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \lambda\,\mathcal{L}_{\text{CE}} \tag{13}$$

where λ is a weighting hyperparameter (commonly set to 1). Dice loss operates directly on region overlap and is comparatively insensitive to the background/foreground imbalance that would otherwise dominate a cross-entropy-only objective, which is pronounced given BRISC's inclusion of non-tumorous and small-lesion cases.

IV. EVALUATION METRICS

At the pixel level, each prediction is scored against ground truth as a true positive (TP), false positive (FP), true negative (TN), or false negative (FN). From these, precision and recall are calculated:

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN} \tag{14}$$

Precision penalizes over-segmentation (false positives) — directly relevant here, since the entire point of the prior-informed module is to suppress anatomically implausible false positives — while recall penalizes under-segmentation (missed tumor regions). Overlap is then summarized by the Dice coefficient [6] and Intersection-over-Union (Jaccard index) [7], computed using binarized predictions:

$$\text{Dice} = \frac{2TP}{2TP+FP+FN}, \qquad \text{IoU} = \frac{TP}{TP+FP+FN} \tag{15}$$

Dice can equivalently be written as the harmonic mean of precision (P) and recall (R), and Dice and IoU are not independent — each determines the other:

$$\text{Dice} = \frac{2\,P\cdot R}{P+R}, \quad \text{IoU} = \frac{P\cdot R}{P+R-P\cdot R}, \quad \text{Dice} = \frac{2\,\text{IoU}}{1+\text{IoU}}, \quad \text{IoU} = \frac{\text{Dice}}{2-\text{Dice}} \tag{16}$$

Both are area-overlap measures and can look favorable even when a model places small false-positive blobs far from the true lesion, since a few misplaced pixels barely move an otherwise-large overlap ratio. Because this is exactly the failure mode the prior-informed module targets, we additionally report a boundary-sensitive metric, the Hausdorff distance [8] between the ground-truth boundary G and the predicted boundary Ŷ:

$$d_H(G, \hat{Y}) = \max\left\{ \sup_{g \in G} \inf_{y \in \hat{Y}} \|g - y\|,\ \sup_{y \in \hat{Y}} \inf_{g \in G} \|g - y\| \right\} \tag{17}$$

A single spatially-isolated false positive — exactly the pattern anatomical priors are meant to suppress — can dominate d_H even when the rest of the boundary is accurately delineated (e.g. a prediction with 99% boundary agreement at 1 px but one stray region 50 px from the lesion yields $d_H = 50$ px). We therefore also report HD95, the 95th percentile of the directed boundary-distance distribution in each direction, with the larger of the two taken as the final score; this preserves sensitivity to boundary disagreement while discounting single extreme outliers. Metrics are reported per class (glioma, meningioma, pituitary) and as a class-weighted average to account for BRISC's class imbalance.

V. EXPERIMENTAL PROTOCOL

Weights are updated by gradient descent using the Adam optimizer [9] with an initial learning rate of $\eta = 1\times10^{-3}$, exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the first and second moment estimates, a numerical stability constant $\varepsilon = 10^{-8}$, and no weight decay. Adam adapts the effective learning rate per parameter from running estimates of the first and second moments of the gradient, giving faster, more stable convergence than plain SGD at the cost of higher memory overhead. The learning rate is decayed with a cosine annealing schedule [10] over 15 epochs, and training uses mini-batches of 16 slices.

To control overfitting — where training performance keeps improving while validation performance degrades, i.e. the model memorizes training-set noise rather than learning transferable structure — we apply dropout [11] within the encoder/decoder blocks, batch normalization [12], and on-the-fly data augmentation (random flips, small rotations of up to ±10°, and intensity jitter on each training slice). Geometric transformations are applied identically to the image, the mask and, for the proposed model, the prior map, so that the prior stays aligned with the image. Conversely, if validation performance plateaus early alongside flat training loss (underfitting), model capacity or training epochs would need to be increased; this is assessed from training/validation loss curves monitored throughout training.

Model selection uses K-fold cross-validation on the training split with K = 3. The training images are randomly shuffled (fixed seed 42) and partitioned into three approximately equal folds. Each model is trained on two folds and validated on the held-out fold, and this is repeated until every fold has served once as the validation fold, giving three trained models per architecture. Within each run, the validation Dice is computed after every epoch and the weights from the epoch with the highest validation Dice are retained, rather than the final-epoch weights. This acts as checkpoint-based model selection that guards against overfitting in later epochs. Each of the three retained models is then evaluated on BRISC's held-out test set, and we report the mean ± standard deviation across the three folds. This gives a more robust estimate of generalization than a single train/validation split.

VI. ABLATION STUDIES

To isolate the contribution of the prior-informed attention module, we compare: (i) the vanilla U-Net with no attention mechanism; (ii) the U-Net with the same gating architecture (Eqs. 9–10) but with $\tilde{P}$ replaced by a learned, data-driven attention map (no anatomical prior); and (iii) the full prior-informed U-Net (proposed). This isolates gains attributable to attention/gating in general from those specifically attributable to injecting the anatomical prior.

VII. RESULTS

Results are reported on the held-out BRISC test set for the three models of Section II-E: the vanilla U-Net (baseline), the U-Net with a learned gate (ablation), and the U-Net with the prior-informed gate (proposed). Model training is described in Section VII-A, overall accuracy in Table I (Section VII-B), false-positive behavior in Fig. 4 (Section VII-C), the ablation in Fig. 5 (Section VII-D), and example predictions in Fig. 6 (Section VII-E); Section VII-F lists the main findings.

A. Model Training

The training split contains 3,933 images and the held-out test set 860 (254 glioma, 306 meningioma, 300 pituitary). Fig. 3 shows training loss, validation loss and validation Dice across the three folds. The three U-Net variants converge almost identically and show no sign of overfitting: training and validation loss decrease together, and validation Dice is still rising slowly at epoch 15, so the models are, if anything, slightly under-trained.

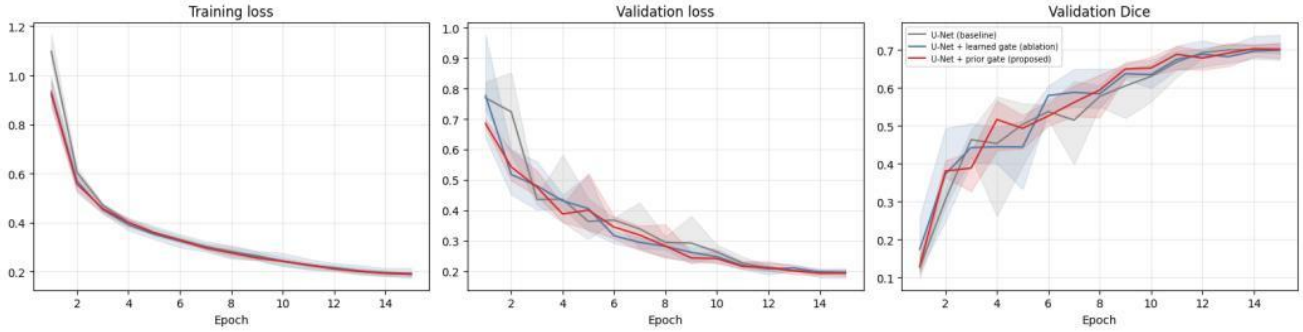


Fig. 3. Training dynamics of the three U-Net variants: training loss (left), validation loss (middle) and validation Dice (right) versus epoch. Lines show the mean over the three cross-validation folds and shaded bands the minimum–maximum range.

B. Quantitative Comparison

Table I reports test-set performance as the mean ± standard deviation over the three cross-validation fold models, with HD, HD95 and false-positive counts in pixels at 224×224 resolution.

TABLE I. Test-set results of the three models (mean ± std over the three cross-validation fold models). HD, HD95 and false-positive counts are in pixels at 224×224; far false positives are false-positive pixels lying more than 20 px from the true lesion; per-class Dice is the mean over the test images of each class. ↑ higher is better, ↓ lower is better; best value per row in bold.

Metric	U-Net (baseline)	U-Net + learned gate (ablation)	U-Net + prior gate (proposed)
Model size			
Parameters (M)	1.93	1.93	1.93
Overlap accuracy			
Dice ↑	0.708 ± 0.012	0.705 ± 0.025	**0.709 ± 0.007**
IoU ↑	0.616 ± 0.011	0.614 ± 0.024	**0.617 ± 0.008**
Precision ↑	**0.778 ± 0.006**	0.768 ± 0.018	0.776 ± 0.016
Recall ↑	**0.721 ± 0.023**	0.719 ± 0.026	0.719 ± 0.009
Boundary accuracy (px)			
HD ↓	40.182 ± 1.818	41.760 ± 6.381	**38.926 ± 1.457**
HD95 ↓	37.858 ± 1.831	38.957 ± 6.582	**36.495 ± 1.509**
False positives (pixels per image)			
All false positives ↓	**134.184 ± 17.617**	147.617 ± 13.453	139.202 ± 14.556
Far false positives ↓	**47.962 ± 11.919**	57.771 ± 10.395	49.728 ± 2.689
Per-class Dice ↑			
Glioma	0.503	0.501	**0.522**
Meningioma	**0.809**	0.806	0.802
Pituitary	**0.779**	0.774	0.773

All three models have the same size (1.93 M parameters) and reach almost identical overlap accuracy. The proposed prior-gated U-Net achieves Dice 0.709 and IoU 0.617, versus 0.708 and 0.616 for the vanilla U-Net and 0.705 and 0.614 for the learned-gate variant, differences far smaller than the fold-to-fold variation (Dice standard deviation 0.007–0.025). The Dice difference between the proposed model and the vanilla U-Net is not significant (paired Wilcoxon signed-rank test [13] over the 860 test images, $p = 0.66$). Precision is marginally lower for the proposed

model (0.776 versus 0.778, p = 0.012) and recall is essentially unchanged (0.719 versus 0.721); the precision difference is statistically detectable but practically negligible. Boundary metrics favor the proposed model slightly: HD is 38.9 px versus 40.2 px (about 3% lower) and HD95 is 36.5 px versus 37.9 px (about 4% lower), but the HD95 difference is not significant (p = 0.26). The proposed model is also the most stable across folds, with the smallest standard deviation in Dice and in both Hausdorff metrics, whereas the learned-gate variant is the least stable (HD95 standard deviation 6.6 px versus 1.5 px).

The false-positive rows of Table I do not show the reduction that the prior was designed to provide. The proposed model produces 139.2 false-positive pixels per image versus 134.2 for the vanilla U-Net, and 49.7 far false-positive pixels versus 48.0 (p = 0.82); the learned-gate variant is higher on both counts (147.6 and 57.8). These measures are analyzed further in Section VII-C. Per class, all three models reach Dice of about 0.80 on meningioma (0.802–0.809) and 0.77–0.78 on pituitary tumors (0.773–0.779), while glioma is by far the hardest class (0.501–0.522). The proposed model is highest on glioma (0.522 versus 0.503 for the vanilla U-Net) but slightly lower on meningioma and pituitary tumors; these per-class differences are at most 0.02, were not tested for significance, and should be interpreted cautiously given the fold-to-fold variation.

C. False-Positive Analysis

Because the central hypothesis concerns anatomically implausible false positives, Fig. 4 reports total false-positive pixels per image, false-positive pixels lying more than 20 px from the true lesion (“far” false positives), and the fraction of images with at least 25 such pixels. The prior gate did not reduce any of these: far false positives averaged 49.7 pixels versus 48.0 for the vanilla U-Net (p = 0.82), and about 14–15% of test images contained a far false positive for all three U-Net variants.

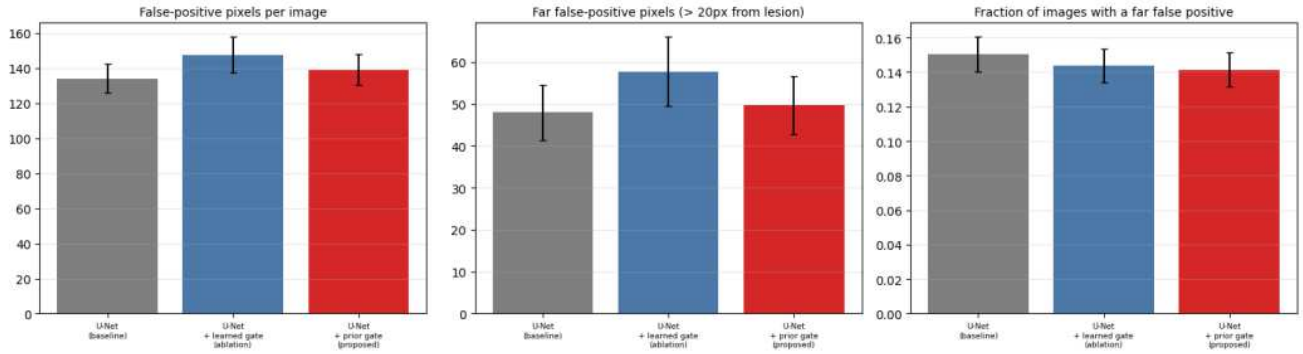


Fig. 4. False-positive analysis on the test set (mean ± standard error over test images): false-positive pixels per image (left), far false-positive pixels lying more than 20 px from the true lesion (middle), and fraction of images containing at least 25 far false-positive pixels (right).

D. Ablation

Fig. 5 isolates the contribution of the gate (Section VI). Replacing the prior-conditioned gate with a learned gate of identical form yielded no gain in Dice or precision and slightly more far false positives (57.8 ± 10.4 versus 49.7 ± 2.7), although this difference lies within the observed variability. We therefore find no evidence that the anatomical prior improves on an ungated U-Net, nor that gating alone helps.

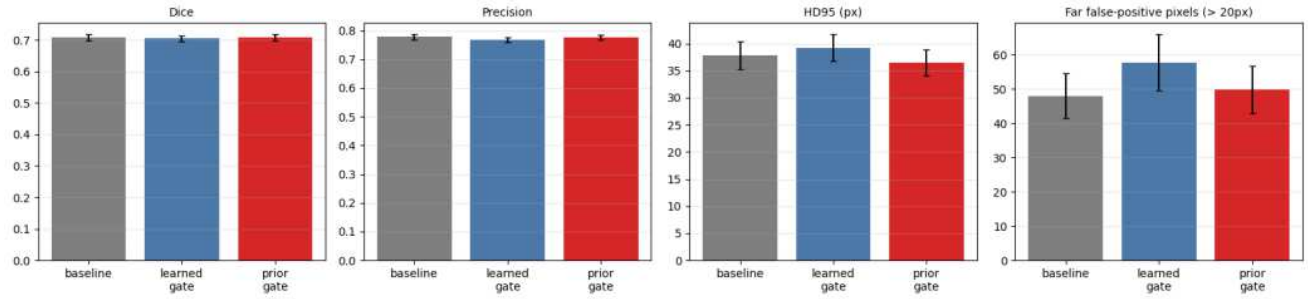


Fig. 5. Ablation of the attention gate on the test set (mean ± standard error): Dice, precision, HD95 (pixels) and far false-positive pixels for the vanilla U-Net (baseline), the U-Net with a learned gate and no anatomical prior, and the U-Net with the prior-informed gate (proposed).

E. Qualitative Results

Fig. 6 shows three randomly chosen test slices and three slices selected for the largest false-positive reduction by the prior gate relative to the vanilla U-Net. In the random examples, all U-Net variants localize the lesion but show thin boundary false positives around small pituitary tumors, and a small low-contrast lesion is missed by every model. In the selected cases, the vanilla U-Net places spurious regions in the frontal lobes or near the skull base, the learned-gate variant shows comparable or larger spurious regions, and the prior gate removes most of them. Because these cases were chosen for the largest improvement, they illustrate the intended mechanism but not typical behavior: across the full test set the effect is not significant (Section VII-C).

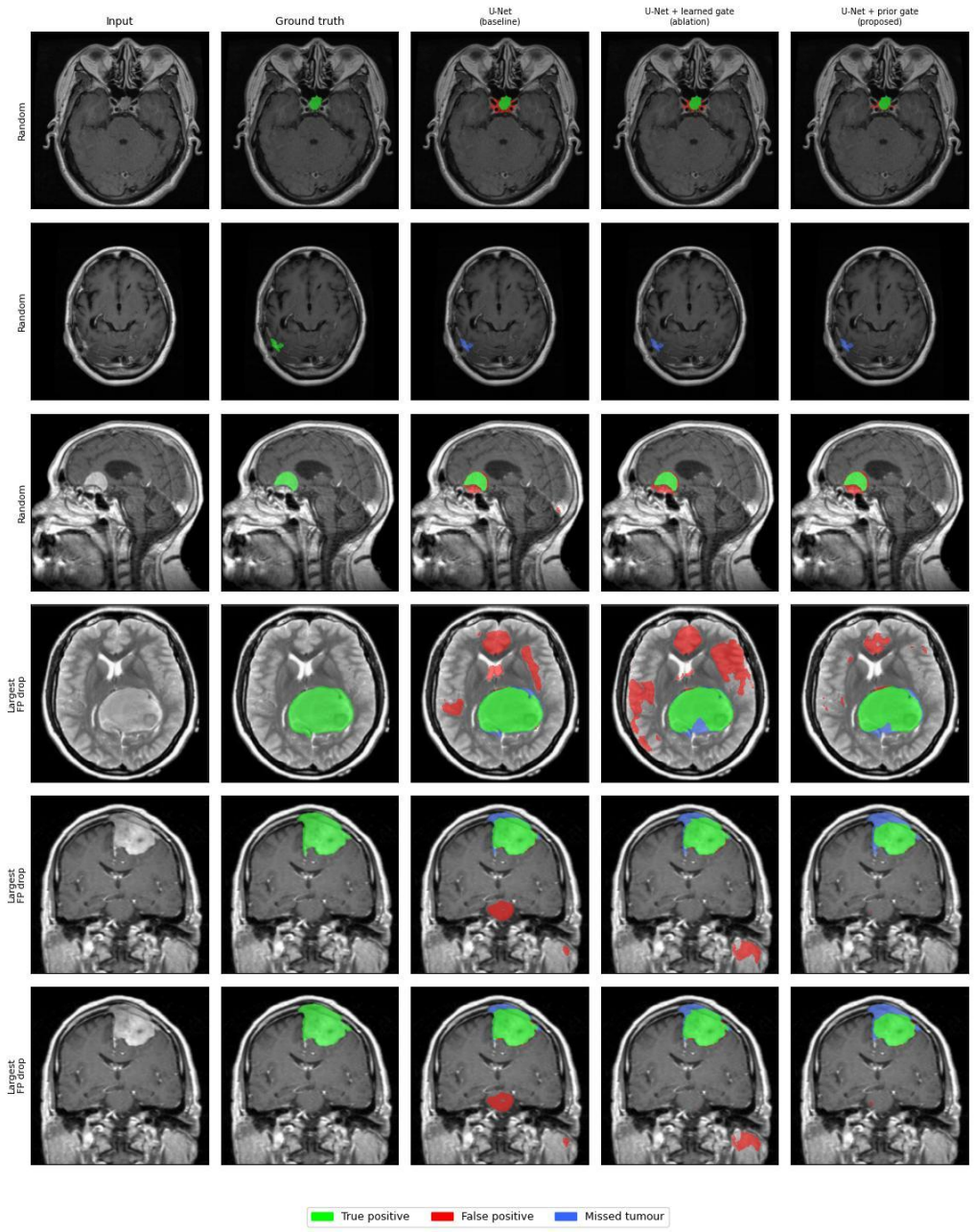


Fig. 6. Qualitative test-set predictions. Columns show the input slice, the ground truth, and the predictions of the vanilla U-Net, the learned-gate U-Net and the proposed prior-gate U-Net. Green: true positive; red: false positive; blue: missed tumor. Rows 1–3 are randomly selected slices; rows 4–6 are the slices with the largest reduction in far false-positive pixels from the vanilla U-Net to the proposed model.

F. Summary of Findings

Overall, the prior-informed gate did not change segmentation performance relative to the vanilla U-Net. Overlap accuracy was unchanged (Dice 0.709 versus 0.708, IoU 0.617 versus 0.616; Dice $p = 0.66$), and boundary accuracy was similar (HD95 36.5 versus 37.9 px, $p = 0.26$, not significant). False positives were not reduced: far false-positive pixels averaged 49.7 versus 48.0 ($p = 0.82$), and about 14–15% of test images contained a far false positive for all three models. The ablation showed no benefit from gating itself, as the learned gate did not improve Dice or precision and yielded slightly more far false positives than the prior gate, within the observed variability. Across all models, glioma

was the most difficult class (Dice 0.50–0.52), compared with about 0.80 for meningioma and 0.77–0.78 for pituitary tumors.

VIII. CONCLUSION

U-Net-based brain tumor segmentation can produce false positives in anatomically implausible locations, which inflate tumor volume estimates and reduce clinical trust. We proposed a prior-informed attention gate that conditions U-Net features on empirical tumor-location priors from the BRISC training annotations, and compared it with a vanilla U-Net and a learned-gate ablation using 3-fold cross-validation. The proposed model matched the baseline in Dice (0.709 versus 0.708) and did not significantly reduce far false positives (49.7 versus 48.0 px, $p = 0.82$). A simple location prior injected through a spatial gate is therefore insufficient on BRISC; stronger prior integration, longer training over multiple seeds, and additional datasets are needed to establish any clinical benefit.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241, Springer, 2015.

[2] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.

[3] M. Ghaffari, A. Sowmya, and R. Oliver, "Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the BraTS 2012–2018 challenges," IEEE Reviews in Biomedical Engineering, vol. 13, pp. 156–168, 2019.

[4] A. Fateh et al., "BRISC: Annotated dataset for brain tumor segmentation and classification," Scientific Data, vol. 13, p. 361, 2026.

[5] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 240–248, Springer, 2017.

[6] L. R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.

[7] P. Jaccard, "The distribution of the flora in the alpine zone," New Phytologist, vol. 11, no. 2, pp. 37–50, 1912.

[8] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 9, pp. 850–863, 1993.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), 2015.

[10] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in International Conference on Learning Representations (ICLR), 2017.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, no. 56, pp. 1929–1958, 2014.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of the 32nd International Conference on Machine Learning (ICML), pp. 448–456, 2015.

[13] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945.