Applied Soft Computing Journal 197 (2026) 115252

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Uncertainty-aware semi-supervised learning for neurosurgical navigation

Francesco Nitti [a], Silvia Seoni [a], Alberto Morello [b], Lorenzo Dolci [b], Amedeo Piazza [c], Vincenzo Esposito [c], Luigi Rosito [c], Diego Garbossa [b], Fabio Cofano [b], Abdulkadir Sengur [d], Massimo Salvi [a,*]

[a] Biolab, PoliToBIOMed Lab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin 10129, Italy
[b] Neurosurgery Unit, Department of Neuroscience "Rita Levi Montalcini", AOU Città della Salute e della Scienza di Torino, University Hospital, University of Turin, Turin 10126, Italy
[c] Neurosurgery Unit, Sapienza University of Rome, Rome, Italy
[d] Electrical-Electronics Engineering Department, Technology Faculty, Firat University, Elazig, Turkey

HIGHLIGHTS

- Novel Spatial Uncertainty metric guides semi-supervised neurosurgical segmentation
- Class-specific calibration ensures reliable pseudo-labeling with adaptive thresholds
- Major gains for rare classes (+40% Dice for tumors, +14% for aneurysms)
- Outperforms data augmentation and generalizes across different architectures
- Real-time speed (3ms/frame) enables practical intraoperative use

ARTICLE INFO

Keywords:
Semi-supervised learning
Uncertainty quantification
Neurosurgical navigation
Pseudo-labeling
Monte Carlo Dropout

ABSTRACT

Background/Objective: Accurate, real-time segmentation of anatomical structures during neurosurgical procedures can support intraoperative orientation. One of the most significant challenges in this domain is developing robust segmentation models with limited annotated data while maintaining clinical reliability. This work addresses how semi-supervised learning can leverage both labeled and unlabeled data, while ensuring the dependability crucial for clinical applications where even small segmentation errors can have significant consequences. *Methods:* We present a novel uncertainty-aware semi-supervised framework for neurosurgical scene segmentation. Our approach introduces Semantic Spatial Uncertainty (SSU), a metric that quantifies prediction reliability by analyzing spatial consistency across multiple stochastic forward passes using Monte Carlo Dropout. The framework employs class-specific calibration with adaptive thresholds that continuously refine through iterative pseudo-labeling, effectively counteracting dataset imbalance. *Results:* Our method achieves significant improvements for clinically critical classes, with relative gains in Dice Similarity Coefficient of + 40% for tumors, + 15% for middle cerebral artery and + 14% for aneurysm. Unlike traditional uncertainty measures, SSU captures uncertainty even for structures with high perimeter-to-area ratios, demonstrating strong correlation with segmentation quality (Pearson coefficient −0.85) without requiring ground truth. Our approach also outperforms intensive data augmentation (even at 200% synthetic samples) and maintains effectiveness across multiple architectures, demonstrating its architecture-agnostic advantages. *Conclusion:* By reframing annotation scarcity as an uncertainty quantification problem, our approach provides a practical solution for medical image segmentation in data-constrained environments. This segmentation framework offers potential applications beyond neurosurgery to other computer vision segmentation tasks with limited labeled data. Code is available at https://github.com/nittifra/ua-ssl-neuro

* Correspondence to: Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi, Turin 24 – 10129, Italy.
E-mail addresses: francesco.nitti@polito.it (F. Nitti), silvia.seoni@polito.it (S. Seoni), alberto.morello@unito.it (A. Morello), lorenzo.dolci@unito.it (L. Dolci), amedeo.piazza@icloud.com (A. Piazza), vincenzo.esposito@uniroma1.it (V. Esposito), luigi.rosito@uniroma1.it (L. Rosito), diego.garbossa@unito.it (D. Garbossa), fabio.cofano@unito.it (F. Cofano), ksengur@firat.edu.tr (A. Sengur), massimo.salvi@polito.it (M. Salvi).

https://doi.org/10.1016/j.asoc.2026.115252
Received 28 June 2025; Received in revised form 24 December 2025; Accepted 7 April 2026
Available online 15 April 2026

1. Introduction

Due to the complexity of neuroanatomy, neurosurgical navigation plays an increasingly important role in modern neurosurgery, offering real-time support to surgeons during cerebral procedures [1]. The application of computer vision algorithms relates to real-time anatomical recognition, based on live image feedback from the surgical device, has the potential to serve as a reliable tool for intraoperative orientation. Unlike pre-operative imaging techniques such as MRI, which offer static anatomical information, intraoperative neurosurgical navigation must contend with dynamic tissue changes that occur during surgery [2,3]. These changes, often referred to as brain shift, arise from factors like cerebrospinal fluid loss, surgical manipulation, and gravitational effects, making accurate and reliable real-time anatomical identification essential for patient safety and surgical success [4].

The identification and segmentation of critical structures during neurosurgery presents significant challenges [5]. Anatomical structures often lack clear boundaries, especially in the presence of bleeding or surgical manipulation. Additionally, surgical scenes frequently involve subtle structures, such as arterial branches or aneurysms, that may only become visible partway through a procedure, requiring continuous adaptation of the navigation system [6]. The complexity is further increased by low-contrast imagery, non-rigid tissue deformation, and visual disruptions caused by surgical tools and fluids, making accurate segmentation challenging even for experienced surgeons [7].

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have enabled more accurate real-time segmentation of surgical scenes [8]. For instance, U-Net architectures have demonstrated success in identifying tumors, vessels, and neural tissues in medical imaging [9]. However, these approaches typically require large quantities of densely annotated training data, which is particularly challenging to obtain in neurosurgical contexts due to the time-consuming nature of per-pixel annotations and the relative scarcity of cases [10].

Recently, semi-supervised learning has emerged as a promising paradigm to address the limitations of labeled data availability in medical imaging [11–13]. Semi-supervised approaches leverage both labeled and unlabeled data to improve segmentation performance while reducing the dependency on extensive manual annotations [14]. Common techniques in this domain include pseudo-labeling, where the model's predictions on unlabeled data are incorporated into the training process, and consistency regularization, which encourages stable predictions under different perturbations of the input [15,16]. While most approaches focus on architectural innovations, we argue that the key challenge lies in developing reliable estimators that can effectively guide the learning process when working with unlabeled data.

The field has seen significant advances in recent years, with several innovative approaches pushing the boundaries of what's possible with limited labeled data. Zhu et al. [17] introduced Contrastive Cluster Assignment (CCA), demonstrating how cross-attention between pixel features and class features can enhance segmentation performance across both supervised and semi-supervised settings. Zeng et al. [18] proposed VerSemi, a versatile framework that integrates multiple semi-supervised tasks into a unified model, achieving state-of-the-art performance across diverse medical imaging datasets. Recent work has also explored density-based approaches that leverage the geometric properties of feature space, combining label-guided co-training with density-guided geometric regularization to provide more robust supervision for unlabeled data [19].

Particularly relevant to our work is the growing adoption of advanced architectures and novel learning strategies. Yang et al. [20] demonstrated the significant impact of using modern Vision Transformer architectures as backbones for semi-supervised segmentation, showing how pre-training on massive datasets can substantially improve performance compared to traditional approaches. More recently, Chen et al. [21] introduced semi-supervised medical image Segmentation via Diffusion Models (MCSD), leveraging diffusion models to enhance feature extraction and implementing multi-consistency learning strategies, demonstrating how generative models can contribute to more robust semi-supervised segmentation frameworks.

While these approaches demonstrate the potential of semi-supervised learning, they often rely on specific architectural choices or complex training strategies that may not generalize well across different scenarios. Moreover, challenges remain in developing reliable methods that can effectively guide the semi-supervised learning process. The application of semi-supervised learning to neurosurgical navigation requires particular attention to structure identification precision and careful management of prediction errors [22].

A fundamental issue in semi-supervised segmentation is model calibration. Modern neural networks often produce poorly calibrated predictions, meaning they tend to be overconfident even when incorrect [23]. This overconfidence can propagate errors during pseudo-labeling, resulting in sub-optimal training. For instance, in neurosurgical image segmentation, overconfident predictions can lead to incorrect pseudo-labels being treated as ground truth, which in turn causes the model to learn inaccurate features. This can be particularly detrimental when segmenting fine anatomical structures, where even small errors can significantly impact the overall quality of the segmentation, especially if the spatial relationship between anatomical structures is compromised by wrong pseudo-labels [24].

In this scenario, uncertainty quantification (UQ) may play a significant role in addressing these challenges, with uncertainty typically categorized as either aleatoric or epistemic. Aleatoric uncertainty arises from inherent noise in the data, such as low-contrast regions in brain images, making it irreducible. In contrast, epistemic uncertainty stems from a model's lack of knowledge and can theoretically be reduced by incorporating more training data [25]. For example, a model may exhibit epistemic uncertainty if it encounters surgical tools or anatomical variations not seen in the training data, requiring additional learning to improve reliability.

In this work, we take a different approach: rather than proposing a new architecture, we develop an architecture-agnostic framework that can enhance the performance of any existing segmentation model through uncertainty-guided pseudo-labeling. This design choice ensures our method remains valuable even as architectures continue to evolve rapidly. Our framework begins with a fully supervised training phase using labeled data and then moves into an adaptive calibration phase to perform pseudo labeling. By leveraging Monte Carlo Dropout [26] to estimate uncertainty, we can predict segmentation quality even when ground truth is unavailable, thus enabling the selection of high-quality pseudo-labels. The main contributions of this work are:

- We introduce a novel metric called "Semantic Spatial Uncertainty" (SSU) that quantifies uncertainty by analyzing spatial consistency across multiple stochastic predictions. SSU demonstrates a high correlation with segmentation quality, providing a reliable mechanism for selecting high-confidence pseudo-labels in the absence of ground truth.
- We develop an adaptive, class-specific calibration system that dynamically establishes uncertainty thresholds through clustering. This approach addresses class imbalance by prioritizing underrepresented structures while maintaining overall segmentation coherence.
- We demonstrate significant performance improvements on neurosurgical segmentation tasks across different architectures (ViT-DeepLab and Swin). Our experiments show that the framework effectively leverages unlabeled data to enhance segmentation accuracy of anatomical structures and surgical instruments. To the best of our knowledge, this is the first semi-supervised learning approach applied to neurosurgical navigation.
- While developed for neurosurgical applications, this approach extends to other vision tasks. The techniques provide a flexible solution

for medical imaging and segmentation challenges with scarce labeled data, enhancing model performance in data-limited scenarios.

This paper is structured as follows: Section 2 provides a comprehensive overview of the proposed method, while Section 3 details the experimental results. Finally, Sections 4 and 5 offer a thorough discussion of the overall work and their implications for advancing reliable AI-assisted neurosurgical navigation.

2. Materials and methods

We propose a training pipeline for semi-supervised segmentation in neurosurgical navigation, combining a fully supervised initial phase with an uncertainty-guided pseudo-labeling strategy. Reliable pseudo-labels are selected based on class-specific uncertainty thresholds, promoting the inclusion of rare and clinically significant structures. The expanded dataset is then used for a final retraining step to improve segmentation performance.

To foster reproducibility and facilitate further research in this area, we have released the complete implementation of our uncertainty-aware semi-supervised framework as open-source code, accessible at https://github.com/nittifra/ua-ssl-neuro

2.1. Dataset

In this work, we utilized a custom dataset of neurosurgical videos provided by "Città della Salute e della Scienza" University Hospital, Turin, Italy. The dataset consists of surgical footage recorded during open pterional craniotomy procedures, capturing a diverse range of anatomical structures and surgical instruments essential for effective neurosurgical navigation. The dataset presents significant challenges for segmentation due to the complexity of neurosurgical scenes and the labor-intensive nature of dense pixel-wise annotation. Although the videos were recorded at 30 frames per second, only 1 frame per second was considered as containing meaningful information, yielding over 23000 frames of interest. Among these, a total of 1147 frames were densely annotated.

For the experiments, the dataset was divided patient-wise into three subsets: 60% for training, 25% for validation, and 15% for testing. The training set was used for model optimization, the validation set for hyperparameter tuning and pseudo-labeling calibration, and the test set exclusively for final performance evaluation.

The annotations, verified by a junior and a senior neurosurgeon (A.M and F. C., respectively), include 18 distinct classes, comprising both anatomical structures and surgical instruments. The anatomical classes include Middle Cerebral Artery (MCA), MCA Branches, Frontal Lobe, Temporal Lobe, Superficial Middle Vein (SMV), Arachnoid membrane, Aneurysm, Dura matter, Second Cranial Nerve (CN II), Anterior Cerebral Artery (ACA), and Tumor. The surgical instrument classes include Scissors, Forceps, Dissector, Suction, Cottonoid, Surgicel, and Clip. Table 1 presents the composition of the dataset, showing the number of frames in each subset and highlighting the substantial number of unlabeled frames available for semi-supervised learning. The complete dataset used in this study, including all annotated and unlabeled frames, is publicly available at https://www.kaggle.com/datasets/artemis90/neurosurgical-navigation-semisupervised-dataset

Table 1
Dataset composition showing the distribution of frames across subsets.

Subset	Annotated frames	Unlabeled frames	Total frames	% Annotated
Training	688	22253	22.942	3%
Validation	287	-	287	100%
Test	172	-	172	100%
Total	1147	22253	23400	5%

The dataset exhibits significant class imbalance, as illustrated in Table 2 and Table 3, which shows the percentage of annotated frames containing each class. Well-represented classes such as Frontal Lobe appear in 89.3% of the annotated frames, while underrepresented classes like Tumor appear in only 2.5% of frames. This imbalance further motivates our semi-supervised approach, which aims to leverage unlabeled data to improve performance on underrepresented classes.

2.2. Proposed semi-supervised framework

2.2.1. Model architecture

Our segmentation framework integrates a Vision Transformer (ViT) backbone [27] with a DeepLabV3 + decoder architecture [28], combining self-attention mechanisms with effective multi-scale feature processing for dense prediction tasks. The backbone consists of a ViT encoder pre-trained on ImageNet [29], which divides input images into fixed-size patches (16 ×16 pixels). These patches are linearly embedded and augmented with positional encodings before being processed through a series of transformer blocks. Each transformer block contains multi-head self-attention layers followed by MLP blocks with GELU activations. This design allows the model to capture long-range dependencies and contextual relationships across the entire surgical field.

For the decoder, we employed the DeepLabV3 + architecture with an Atrous Spatial Pyramid Pooling (ASPP) module to effectively handle multi-scale contextual information. The ASPP module uses dilated convolutions with different rates (6, 12, and 18) to capture objects at various scales without increasing computational complexity. The decoder fuses high-level semantic features from the transformer backbone with low-level spatial features through skip connections, enabling precise boundary localization.

To reduce the risk of overfitting and enable uncertainty estimation at test time, we integrated spatial dropout layers (rate = 0.5) within the decoder pathways. Fig. 1 illustrates the architecture, which combines a ViT encoder with a DeepLabV3 + decoder to produce the final segmentation masks.

Prior to model training, all frames were resized to a uniform resolution of 512 × 512. To mitigate the effects of varying lighting conditions during surgical procedures, we applied histogram equalization [30] to normalize contrast across frames, followed by intensity normalization to scale pixel values to the range [0,1].

During training, we employed data augmentation to address the

Table 2
Anatomical structures in annotated frames.

Class	Number of annotations	Presence (%)	Description
MCA	520	45.4	Major artery supplying blood to the brain
MCA branches	423	36.9	Branches of the main cerebral artery
Frontal lobe	1024	89.3	Region involved in cognitive functions
SMV	584	50.9	Vein located on the brain's surface
Arachnoid	587	51.2	Delicate membrane covering the brain
Temporal lobe	849	74.1	Region processing auditory and language information
Aneurysm	370	32.3	Abnormal bulge in a blood vessel
Dura	450	39.2	Tough membrane covering the brain
CN II	272	23.7	Optic nerve responsible for vision
ACA	222	19.4	Artery supplying the frontal lobe
Tumor	28	2.5	Abnormal tissue growth

Table 3
Instruments presence in annotated frames.

Class	Number of annotations	Presence (%)	Description
Scissors	157	13.7	Surgical instrument used for cutting
Forceps	413	36.0	Surgical tool used for grasping
Dissector	267	23.3	Instrument used to separate tissues
Suction	800	69.8	Device used to remove fluids
Cottonoid	653	56.9	Absorbent pad used during surgery
Surgicel	184	16.0	Hemostatic agent used to control bleeding
Clip	163	14.2	Surgical clip used for vessel occlusion

limited availability of labeled data. The augmentations included geometric transformations (rotations within ± 20 °, elastic deformations with alpha=1.0 and sigma=50.0, random resized cropping with scaling between 0.7 and 1.0 × original size, optical distortions with shift parameters between -0.05 and 0.05, and perspective changes with scale up to 0.05) and photometric transformations (random brightness and contrast adjustments with limits of ± 0.2, hue and saturation shifts up to $\pm 3\%$, RGB channel shifts within $\pm 5\%$ of the value range, motion blur with kernels up to 5 pixels, and simulated ISO noise with variance up to 0.04). The photometric transformations are particularly relevant for simulating the lighting variations typical in surgical environments, where spotlights can create non-uniform lighting conditions. Color shifts also simulate the slight variations in hue due to bleeding or different tissue types visualized during the procedure. Horizontal and vertical flipping were deliberately excluded to preserve anatomical spatial coherence, as the left-right and superior-inferior orientation of brain structures has fundamental clinical significance that should not be altered.

Each input image received a random subset of these transformations, with all geometric transformations applied identically to both the input image and its corresponding segmentation mask to preserve pixel-wise

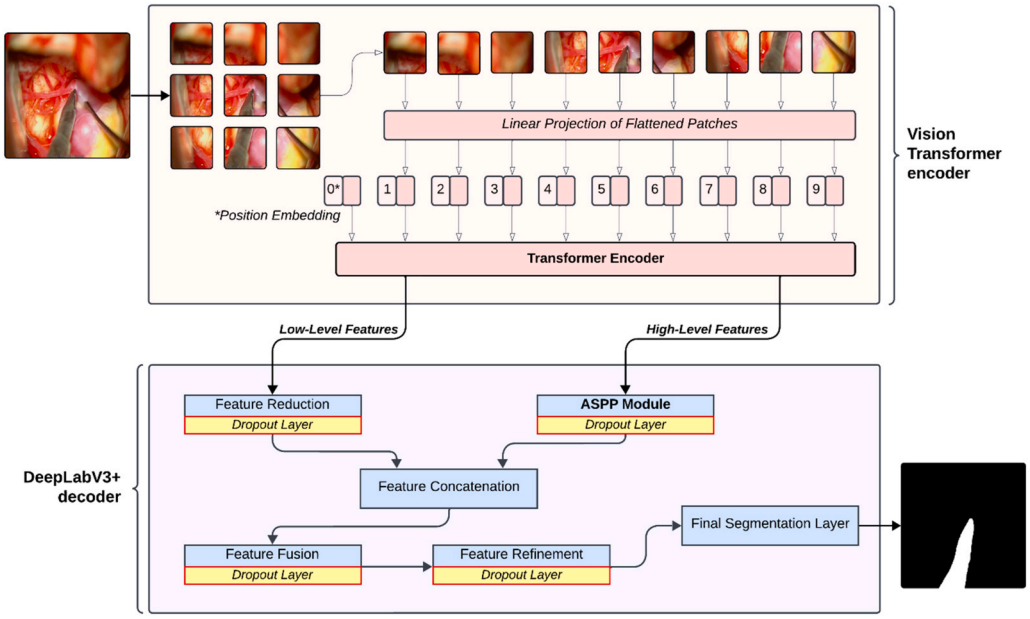

Fig. 1. Overview of the Vision Transformer and DeepLabV3 + used in this work.

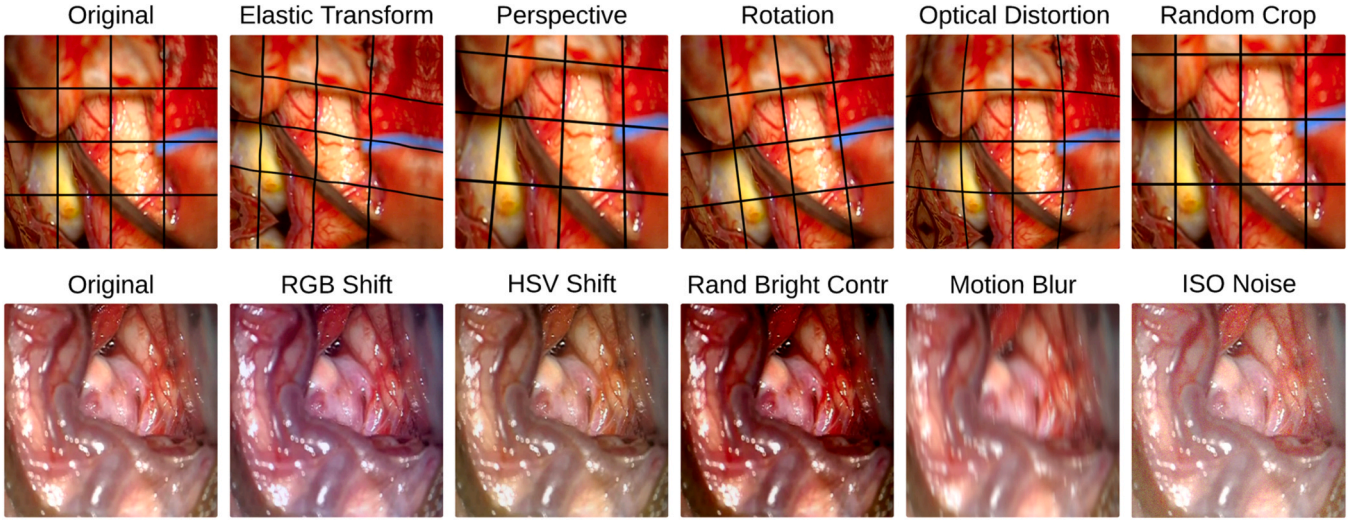

Fig. 2. Examples of augmentations applied to the dataset. Top row illustrates geometric transformations, while bottom row shows color-based augmentations.

correspondence. For unlabeled data in our semi-supervised approach, we applied the same augmentation strategy to ensure feature space consistency. Fig. 2 shows some examples of the augmentations applied to our dataset.

2.2.2. Semantic spatial uncertainty

Traditional UQ methods primarily focus on estimating class uncertainty [31,32], with limited application to segmentation tasks. Even within segmentation, most existing approaches address binary tasks [33,34] failing to recognize that in multi-class scenarios, networks may exhibit varying levels of uncertainty across different classes. To address these limitations, we propose SSU, a novel class-specific uncertainty metric that quantifies spatial disagreement between predictions across multiple stochastic forward passes. Unlike conventional metrics that only consider class probability distributions at each pixel, SSU explicitly accounts for the spatial coherence of segmentation predictions for each semantic class independently.

The idea behind SSU is that uncertainty in medical image segmentation manifests not only as class confusion but also as spatial variance in predicted boundaries. By leveraging Monte Carlo dropout during inference, we generate $K = 25$ segmentation masks for each input image and compute the spatial agreement between these predictions for each class. Mathematically, for a given class, the SSU is defined as:

$$SSU = 1 - \frac{2 \times PRED \cap LOR}{\sum PRED + \sum LOR} \tag{1}$$

where *PRED* represents the mean prediction across K Monte Carlo forward passes, *LOR* (Logical OR) is the union of all positive predictions across these passes, and $\cap$ denotes the intersection operation. The denominator normalizes the metric, ensuring values between 0 and 1, where 0 indicates perfect spatial agreement (no uncertainty) and 1 indicates complete disagreement (maximum uncertainty).

This formulation specifically measures the spatial disagreement between the most sensitive prediction (*LOR*, which captures all pixels predicted as positive in any forward pass) and the average prediction (*PRED*), providing a clear indication of regions where the model exhibits prediction variability. When predictions are stable, *PRED* and *LOR* will be similar in extent, resulting in their intersection being close to their individual areas (low uncertainty). Conversely, in regions where predictions fluctuate between passes, *LOR* will encompass a larger area than *PRED* due to inconsistent predictions, leading to a smaller intersection relative to their total area (high uncertainty). This spatial consistency measure is particularly valuable for surgical navigation, where stable boundary predictions are crucial for clinical reliability.

Fig. 3 shows the application of SSU to a sample frame, specifically for the class 'temporal lobe'. The visualization shows how multiple stochastic predictions through Monte Carlo dropout reveal areas of consistent agreement (low uncertainty) in the central regions of the lobe and areas of higher uncertainty, particularly at the anatomical boundaries where the temporal lobe interfaces with surrounding structures.

2.2.3. Uncertainty-guided pseudo-labeling

One of the main challenges in pseudo-labeling is distinguishing between reliable and unreliable predictions in the absence of ground truth annotations. To address this, we introduce a class-specific calibration strategy based on the joint distribution of model uncertainty and segmentation quality, computed on the validation set where ground truth is available. For each semantic class c, we compute the SSU and Dice Similarity Coefficient (DSC) for all relevant predictions on the validation set, obtaining a set of two-dimensional points $\{(SSU_i, DSC_i)\}_i$. We then apply k-means clustering with $k = 2$ to this set in order to separate predictions into two groups:

- A reliable cluster, typically characterized by low uncertainty (SSU) and high segmentation quality (DSC).
- An unreliable cluster, associated with high uncertainty and low segmentation quality.

We chose $k = 2$ based on the empirical observation that the relationship between SSU and DSC naturally forms two distinct groups, which aligns with our goal of binary classification into reliable versus unreliable predictions. Once the clusters are identified, we extract the SSU coordinate of the reliable cluster's centroid, which we define as the SSU threshold θ_c for class c. This value represents the maximum level of uncertainty that is still associated, on average, with high-quality predictions, as shown in Fig. 4. Formally, let μ_1^c and μ_2^c denote the SSU coordinates of the two clusters centroids obtained for class c, with $\mu_1^c < \mu_2^c$. Then:

$$\theta_c = \mu_1^c \quad (\textit{threshold for class } c) \tag{2}$$

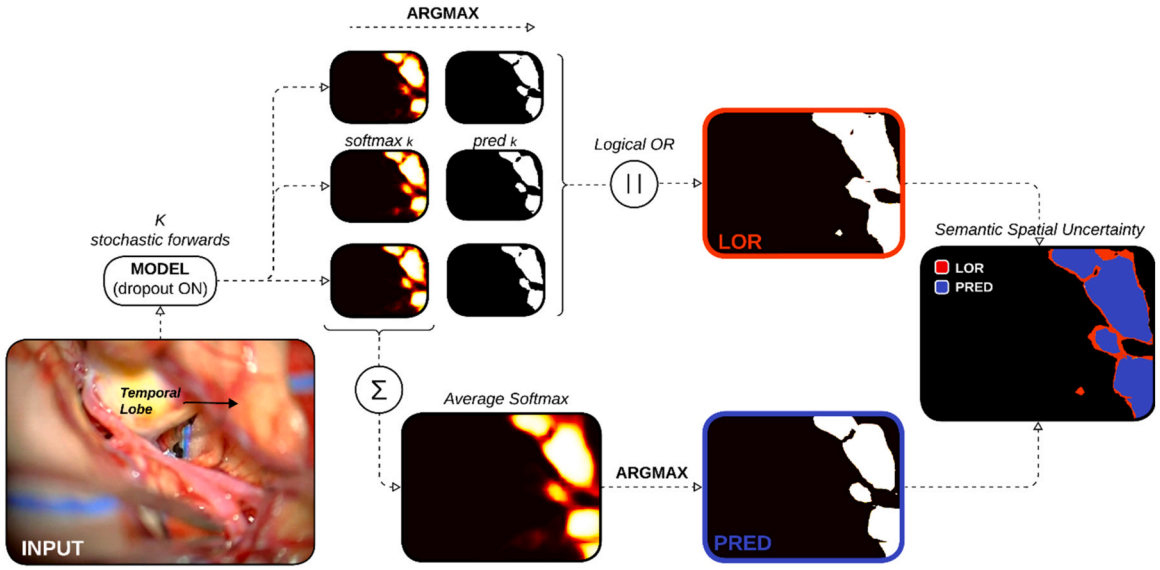

Fig. 3. Computation of Semantic Spatial Uncertainty for a sample class (temporal lobe). The process begins with the input frame, followed by 25 stochastic forward passes to generate the mean prediction (*PRED*) and Logical OR (*LOR*) maps. The resulting SSU map highlights regions of both high (red) and low (blue) spatial uncertainty. The overall SSU value, computed as described in Eq. 1, quantifies the global uncertainty for this class across the entire image.

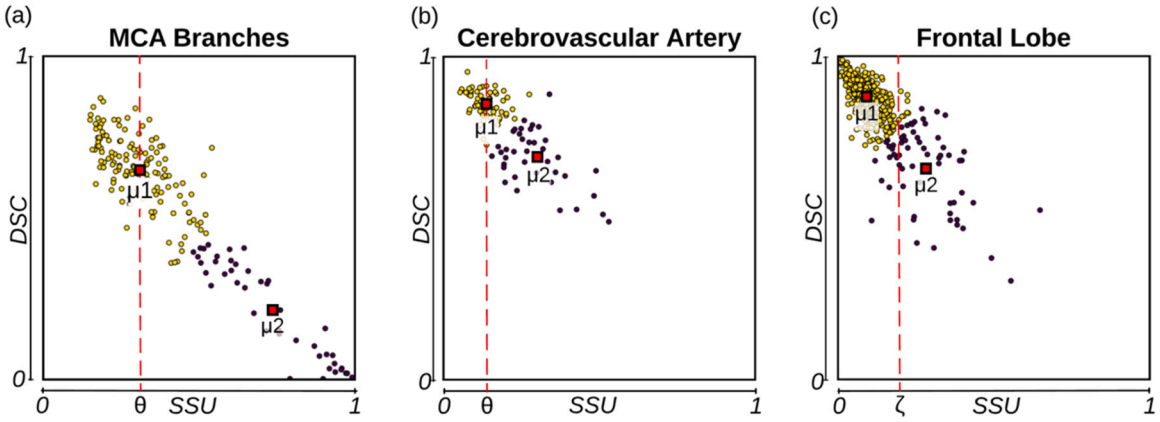

Fig. 4. Uncertainty distribution and adaptive threshold calibration for three representative classes. For each class, SSU values are shown with vertical lines indicating the k-means-derived thresholds that separate reliable predictions (left of the line) from unreliable ones (right of the line). (a) MCA branches (rare class). (b) Anterior Cerebral Artery (moderately frequent class) and (c) Frontal Lobe (common class).

$$\zeta_c = \frac{\mu_1^c + \mu_2^c}{2} \quad (\textit{relaxed threshold for class } c) \tag{3}$$

The standard threshold θ_c is more conservative, representing the center of the reliable cluster, while the relaxed threshold ζ_c represents the midpoint between clusters, allowing for higher uncertainty when appropriate.

As previously shown in Table 2 and Table 3, our dataset presents a significant class imbalance. While some structures appear frequently (e.g., brain tissue, major vessels), others are rare yet clinically critical (e.g., specific cranial nerves, tumor boundaries). This imbalance poses a challenge for standard pseudo-labeling approaches, which tend to reinforce the dominance of common classes while overlooking rare ones [35]. To address this, our pseudo-labeling strategy prioritizes the inclusion of rare classes by using class frequency to determine threshold strictness. We categorize semantic classes into three groups based on their representation in the dataset:

- *Rare classes* (lowest frequency): these classes drive pseudo-label inclusion. A pseudo-label is retained if it contains at least one rare class, and all classes present meet their respective uncertainty criteria. For rare classes, we select the threshold θ_c to ensure only high-confidence predictions are selected. Examples include Forceps, Dissector, Clip, Scissors, Arachnoid, MCA branches, and Tumor.
- *Moderately frequent classes*: these classes are treated with intermediate strictness. The maximum uncertainty threshold θ_c is applied, but their presence is not mandatory to retain a pseudo-label. Examples include Suction, CN II, ACA, MCA, Aneurysm, and SMV.
- *Common classes* (highest frequency): these classes are filtered using the relaxed threshold, ζ_c defined as the midpoint between the two cluster centroids. This allows flexibility for inclusion when rare and intermediate classes are also present, avoiding the exclusion of valuable pseudo-label due to minor uncertainty in well-represented structures. Examples include Frontal lobe, Temporal lobe, Dura, and Cottonoid.

Fig. 4 illustrates examples of threshold estimation for three representative classes spanning our frequency categories. For each class, the scatter plot shows the relationship between SSU and DSC, with the k-means-derived thresholds marked as vertical lines. The threshold position indicates the maximum SSU value considered indicative of a reliable segmentation for that particular class. Note how the threshold placement varies across classes, with stricter (lower) thresholds for rare classes like MCA branches (Fig. 4a) compared to common classes like Frontal Lobe (Fig. 4c).

This threshold assignment strategy effectively "forces" the inclusion of underrepresented classes by allowing predictions to enter the pseudo-labeled set only when rare classes are present, and all included classes meet their respective uncertainty thresholds. In doing so, our method enforces strict reliability for rare structures, while avoiding the over-representation of dominant classes.

After calibrating the uncertainty thresholds on the validation set, the full pseudo-labeling process proceeds as follows:

- The model generates predictions for unlabeled frames using K stochastic forward passes.
- For each semantic class in each frame, the corresponding SSU value is computed.
- Each class is evaluated independently against its assigned uncertainty threshold.
- The pseudo-label is retained only if all classes present in the prediction meet their respective uncertainty criteria.

A visual representation of the pseudo-labeling step is depicted in Fig. 5. This process enriches the training dataset with high-quality predictions generated from the large subset of unannotated frames. As a result, we obtain more data and a more balanced and comprehensive training set, leading to improved segmentation performance across anatomical regions and surgical tools.

2.2.4. Semi-supervised training strategy

Our semi-supervised training approach follows an iterative process designed to progressively leverage both labeled and unlabeled data. As shown in Fig. 6a, the training is structured into the following phases:

- *Initial supervised training*: The model is first trained solely on the manually annotated dataset for 50 epochs. This establishes a performance baseline and enables the learning of core features before introducing pseudo-labels.
- *Adaptive Calibration*: Class-specific uncertainty thresholds are computed on the validation set, as detailed in Section 2.2.3.
- *Pseudo-labeling*: Using the calibrated thresholds, pseudo-labels are generated for the unlabeled data. Only predictions with SSU values below their corresponding class-specific thresholds are retained.
- *Combined training*: The model is retrained for 50 epochs on the augmented dataset composed of both ground-truth labels and high-confidence pseudo-labels.
- *Iterative refinement:* The calibration, pseudo-labeling, and retraining steps are repeated for two additional rounds (three iterations total). Thresholds are recalibrated at each iteration to reflect the model's evolving performance.

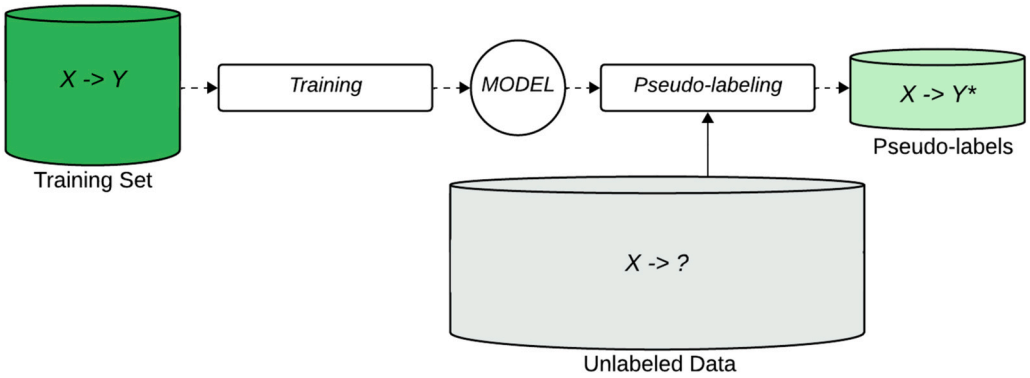

Fig. 5. Overview of the uncertainty-guided pseudo-labeling pipeline. The initial model trained on labeled data ($X \rightarrow Y$) generates predictions ($X \rightarrow Y^*$) for unlabeled frames ($X \rightarrow ?$), which are retained if considered reliable based on the uncertainty criteria.

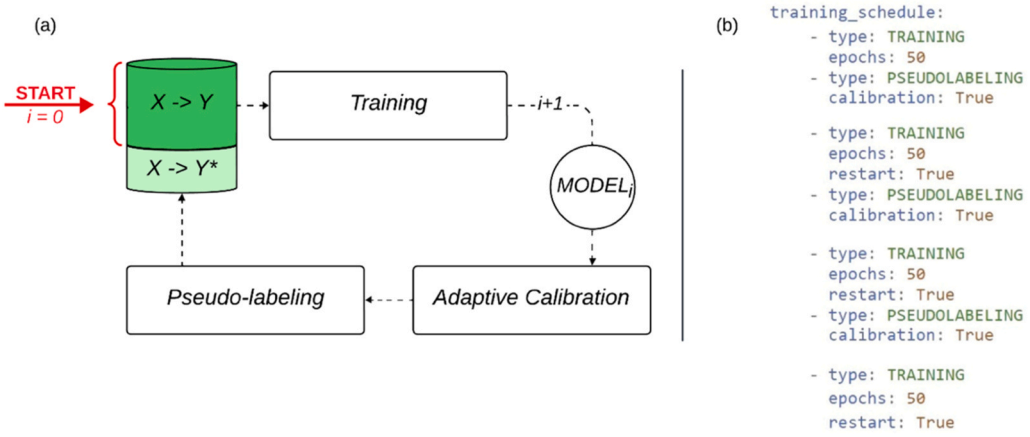

Fig. 6. Semi-supervised training strategy adopted in this work. (a) Workflow of our iterative semi-supervised training process, showing the progression from initial manually labeled training set (X→Y) through multiple pseudo-labeling cycles (X→Y*). (b) Configuration details of our implementation, with calibration performed before each cycle of pseudo-labeling.

Fig. 6b outlines the training configuration used in our implementation. A key strength of this approach is the dynamic, class-specific threshold adjustment at each iteration, which ensures the reliability of pseudo-labels and prevents error accumulation over time. This iterative semi-supervised scheme allows the model to continuously broaden its learning capacity while mitigating confirmation bias.

3. Results

In this section, we present an extensive validation of our proposed semi-supervised framework, including quantitative analyses of uncertainty correlation with segmentation quality, performance enhancements introduced by iterative pseudo-labeling, and evaluations of image-level precision and recall.

3.1. Experimental setup and evaluation protocol

All experiments were conducted on an NVIDIA A100 GPU with 40 GB of memory using PyTorch 2.2.1. For the initial fully supervised training phase, we used Stochastic Gradient Descent with a learning rate of 0.001 (10 × higher for the segmentation head), weight decay of 1e-4, and a cosine annealing scheduler. The Vision Transformer backbone (12 layers, 12 attention heads, 768 hidden dimensions) was initialized with ImageNet-1K pre-trained weights. Training ran for 50 epochs with a batch size of 8. The loss function combined weighted cross-entropy to address class imbalance with a progressively increasing entropy minimization term to encourage prediction confidence. These parameters were optimized for the labeled dataset before initiating the semi-supervised pipeline.

To evaluate segmentation performance, we employed multiple complementary metrics. The Dice Similarity Coefficient (DSC) served as our primary metric for measuring segmentation quality, reported both for individual anatomical structures and as an overall average to capture global performance. We supplemented this with image-level Precision and Recall metrics to assess the model's capability in detecting anatomical structures. Performance was tracked at three key stages throughout the semi-supervised training process: (i) the initial supervised model trained solely on labeled data, (ii) an intermediate checkpoint after the first pseudo-labeling iteration, and (iii) the final model obtained at the end of the proposed pipeline. This longitudinal tracking enabled us to demonstrate the progressive improvements introduced by our approach.

To ensure statistical rigor in our experimental validation, we trained 10 different models with varying random initializations for all ablation experiments and comparisons between different configurations. We performed paired *t*-tests against the baseline with a significance threshold of $p < 0.05$, ensuring that observed improvements were statistically meaningful rather than artifacts of random initialization or training variability. The asterisk (*) notation in tables and figures denotes statistically significant differences compared to the baseline model ($p < 0.05$). Furthermore, we placed particular emphasis on analyzing DSC improvements for underrepresented classes, which typically benefit most from the proposed semi-supervised approach, to validate our method's effectiveness in addressing class imbalance challenges.

3.2. Uncertainty quantification analysis

To evaluate the effectiveness of the proposed SSU as an indicator of segmentation reliability, we analyzed its relationship with segmentation performance across all anatomical structures in our dataset. Fig. 7 illustrates the strong inverse correlation between SSU values and Dice scores, with lower uncertainty values consistently corresponding to higher segmentation performance. Statistical analyses confirmed this relationship, revealing an average Pearson correlation coefficient of −0.85 and an average Spearman correlation coefficient of −0.79, both indicating robust negative correlations across all structures.

We performed a comparative evaluation between SSU and Normalized Entropy (NE), with results summarized in Table 4. Our analysis demonstrates that SSU consistently achieves higher correlation with segmentation quality across all anatomical classes. This improvement is particularly pronounced for vessel-like structures such as MCA branches (SSU: −0.90 vs NE: −0.37 Pearson correlation) and SMV (SSU: −0.85 vs NE: −0.58 Pearson correlation). The superior performance of SSU in these cases can be attributed to its robustness to structure morphology: while NE tends to report high uncertainty for thin structures due to their large perimeter-to-area ratio, SSU maintains reliable uncertainty estimates regardless of structure geometry.

Fig. 8 provides a qualitative comparison through representative cases where SSU and NE produce notably different uncertainty maps.

Table 4
Comparison of correlation coefficients between uncertainty metrics (SSU and NE) and segmentation performance (DSC).

Anatomical structure	Pearson correlation		Spearman correlation	
	SSU	*NE*	*SSU*	*NE*
MCA	-0.82	-0.65	-0.71	-0.72
MCA branches	-0.90	-0.37	-0.85	-0.51
Frontal lobe	-0.79	-0.75	-0.75	-0.71
SMV	-0.85	-0.58	-0.75	-0.78
Arachnoid	-0.90	-0.72	-0.84	-0.73
Temporal lobe	-0.77	-0.76	-0.76	-0.76
Aneurysm	-0.88	-0.71	-0.82	-0.73
Dura	-0.82	-0.75	-0.85	-0.76
CN II	-0.90	-0.73	-0.70	-0.70
ACA	-0.87	-0.67	-0.80	-0.75
Tumor	-0.80	-0.70	-0.71	-0.63

The visual analysis reveals that SSU generates more spatially coherent uncertainty estimates that better align with actual segmentation errors. This is particularly evident in boundary regions and thin anatomical structures, where NE tends to overestimate uncertainty regardless of the actual segmentation quality. The rightmost column of Fig. 8 demonstrates how SSU produces more nuanced uncertainty maps that better reflect the model's true confidence in its predictions.

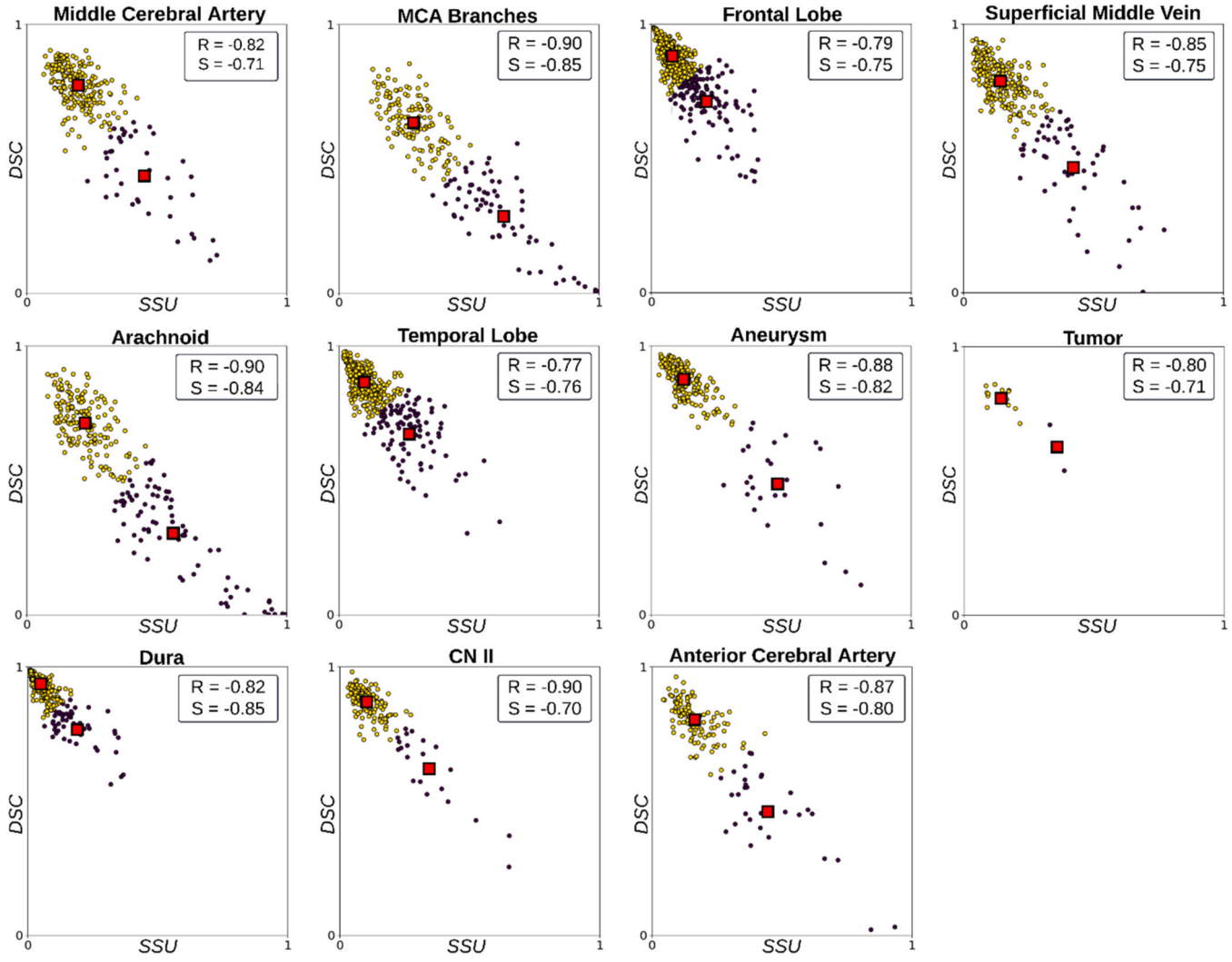

Fig. 7. Scatterplot of Semantic Spatial Uncertainty against Dice Similarity Coefficient for all anatomical classes. R denotes the Pearson correlation coefficient, and S denotes the Spearman correlation coefficient.

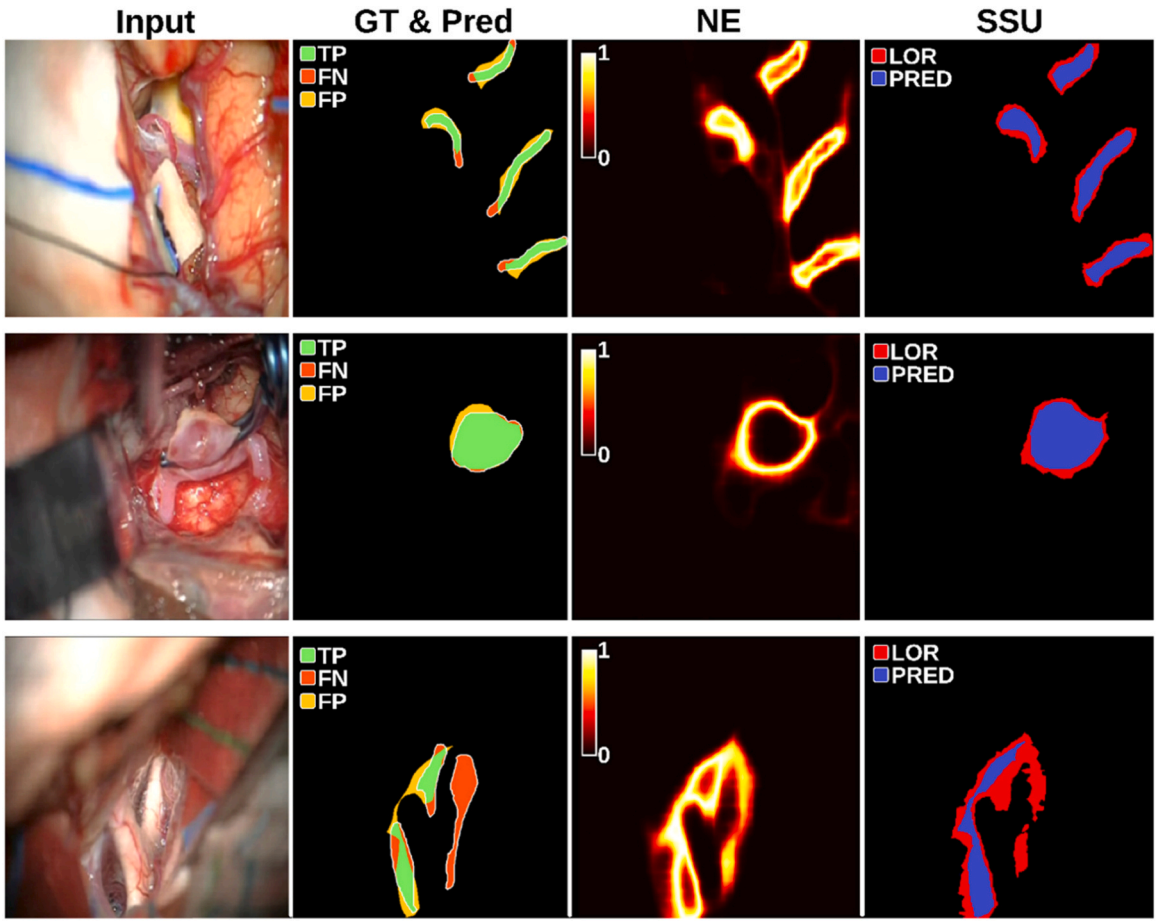

Fig. 8. Each row presents a different neurosurgical scene, showing from left to right: the original input image, ground truth and segmentation overlay, the uncertainty map generated using NE, and the uncertainty map generated using our proposed SSU. Note how SSU (rightmost column) provides more spatially consistent uncertainty estimates that better correlate with actual segmentation errors.

3.3. Segmentation performance analysis

The iterative pseudo-labeling process demonstrated significant performance improvements as shown in Fig. 9. In our experiments, we observed that the first iteration typically adds approximately 2.1% of the unlabeled data, the second iteration adds an additional 1.9%, and the third iteration adds approximately 1.4% more. This diminishing return pattern suggests that the model reaches a point of convergence where the remaining unlabeled data is either too complex or contains anomalies that the model correctly identifies as high uncertainty regions.

The final model demonstrates substantial improvements for several clinically important but underrepresented classes following the full pseudo-labeling process. For instance, the Tumor class showed a + 40% relative increase in DSC, while other rare classes such as Dissector

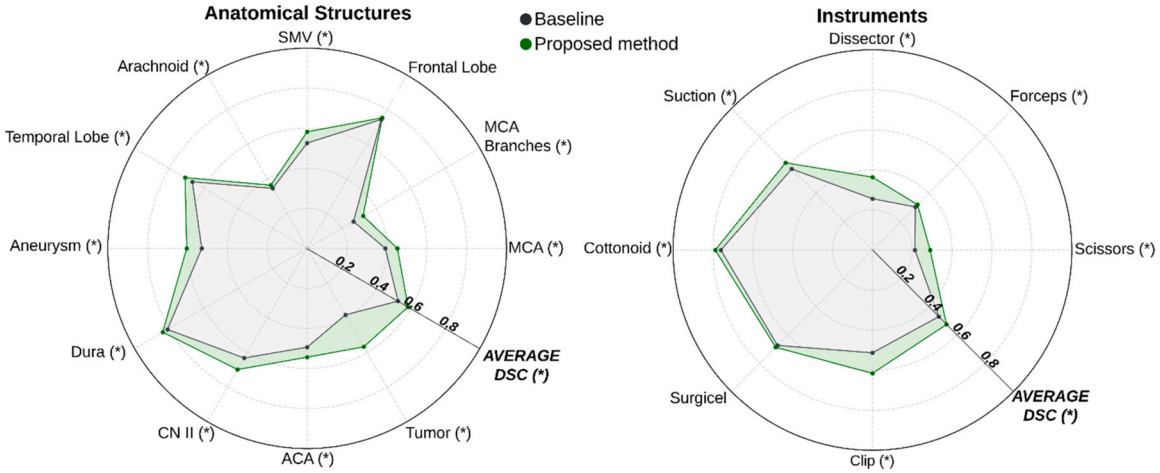

Fig. 9. Effect of the proposed pipeline (three rounds of pseudo-labeling) compared to the baseline. The radial axis represents the Dice Similarity Coefficient. The asterisk (*) denotes classes where the improvement was found to be statistically significant ($p < 0.05$).

(+42%), Scissors (+36%), Clip (+19%), MCA Branches (+21%) and Aneurysm (+14%) also benefited significantly. In contrast, well-represented structures like Frontal Lobe showed only minor gains (approximately +1%), suggesting performance saturation. All reported results represent the average performance over 10 independent runs, with asterisk (*) denoting statistically significant improvements ($p < 0.05$).

Our analysis reveals distinct patterns in image-level detection metrics (Fig. 10): we observe a consistent increase in precision and a decrease in recall across several classes, especially for surgical instruments. This behavior reflects the model's shift toward more conservative and selective predictions after being trained with uncertainty-filtered pseudo-labels. In our image-level evaluation, a class is considered detected if the model predicts even a single pixel belonging to that class. Before pseudo-labeling, the initial model tended to overpredict instruments, frequently misclassifying background elements as tools. After pseudo-labeling, the predictions become more class-specific and spatially coherent, reducing false positives while occasionally leading to missed detections when predictions fall just outside the annotated region. This trade-off is particularly evident for classes such as Clip, Dissector, and Scissors. Despite this, the overall improvement in segmentation quality is confirmed by the consistent gain in Dice score, which better reflects pixel-level accuracy.

Table 5 provides a summary of the overall segmentation performance at each training stage. Metrics reflect the trends already discussed: improvements in Dice and Precision across iterations, alongside a modest drop in Recall due to more selective predictions.

Fig. 11 presents qualitative results for clinically relevant anatomical structures in neurosurgical navigation. These include anatomical regions such as the middle cerebral artery and aneurysms. The visualizations demonstrate the model's ability to produce accurate and coherent segmentations despite common intraoperative challenges like motion blur, partial occlusions, and low tissue contrast. The color-coded overlays and labeled bounding boxes highlight precise delineation of critical regions such as arterial branches and aneurysms.

Table 5
Quantitative comparison of segmentation performance across training stages. The asterisk (*) denotes classes where the difference was found to be significant respect to the baseline ($p < 0.05$).

Model Stage	mDice	mPrecision	mRecall
Baseline (fully supervised)	0.503	0.818	0.914
After 1 iteration of pseudo-labeling	0.531 (*)	0.856 (*)	0.910
Final model (3 iterations of pseudo-labeling)	0.560 (*)	0.879 (*)	0.884 (*)

3.4. Comparative analysis and ablation studies

To validate the versatility of our semi-supervised framework, we evaluated its performance across different architectures and compared it against traditional data augmentation approaches. This analysis aimed to determine whether our iterative pseudo-labeling approach represents a broadly applicable methodology that offers benefits beyond conventional data augmentation techniques, regardless of the underlying network architecture.

Specifically, we implemented our semi-supervised framework on two

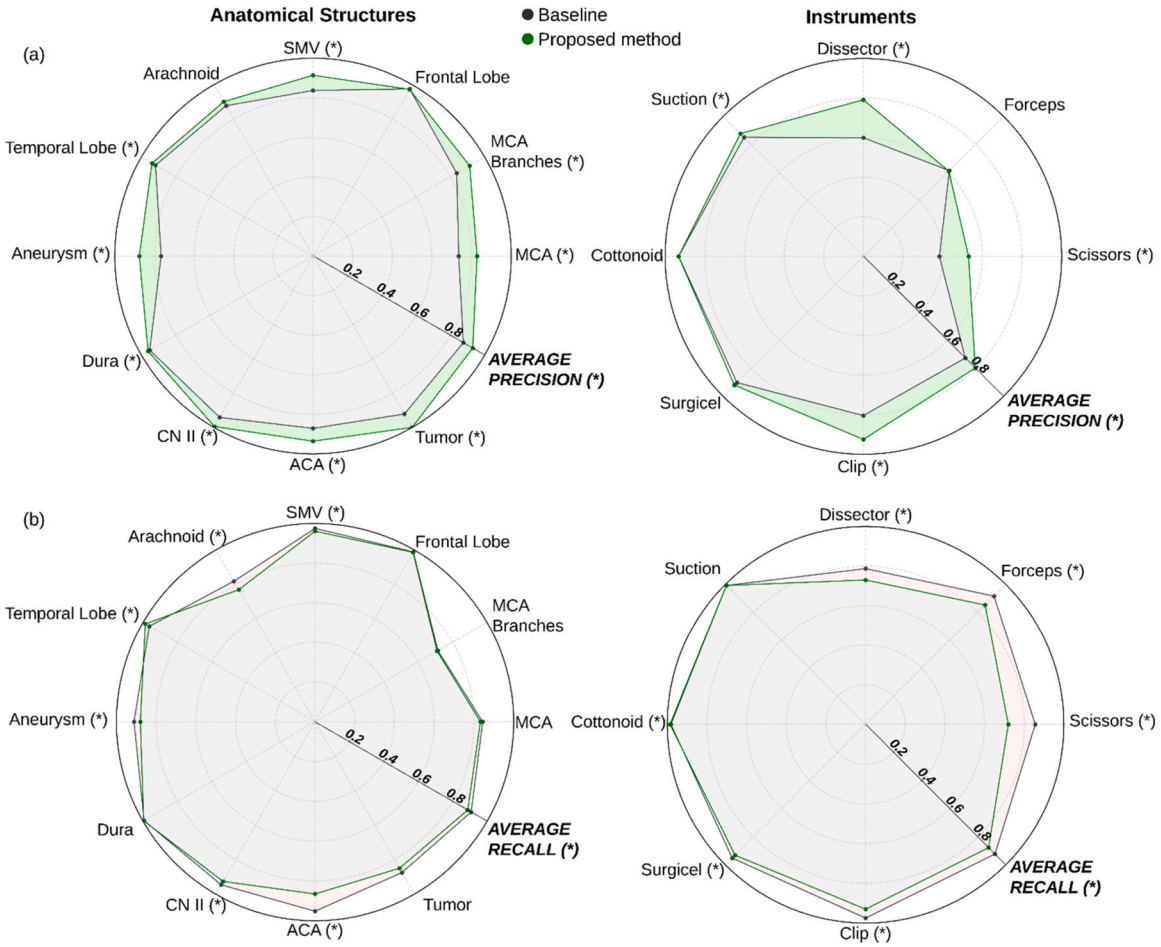

Fig. 10. Spider plots comparing (a) Precision and (b) Recall for anatomical structures and instruments. The asterisk (*) denotes classes where the difference was found to be significant ($p < 0.05$).

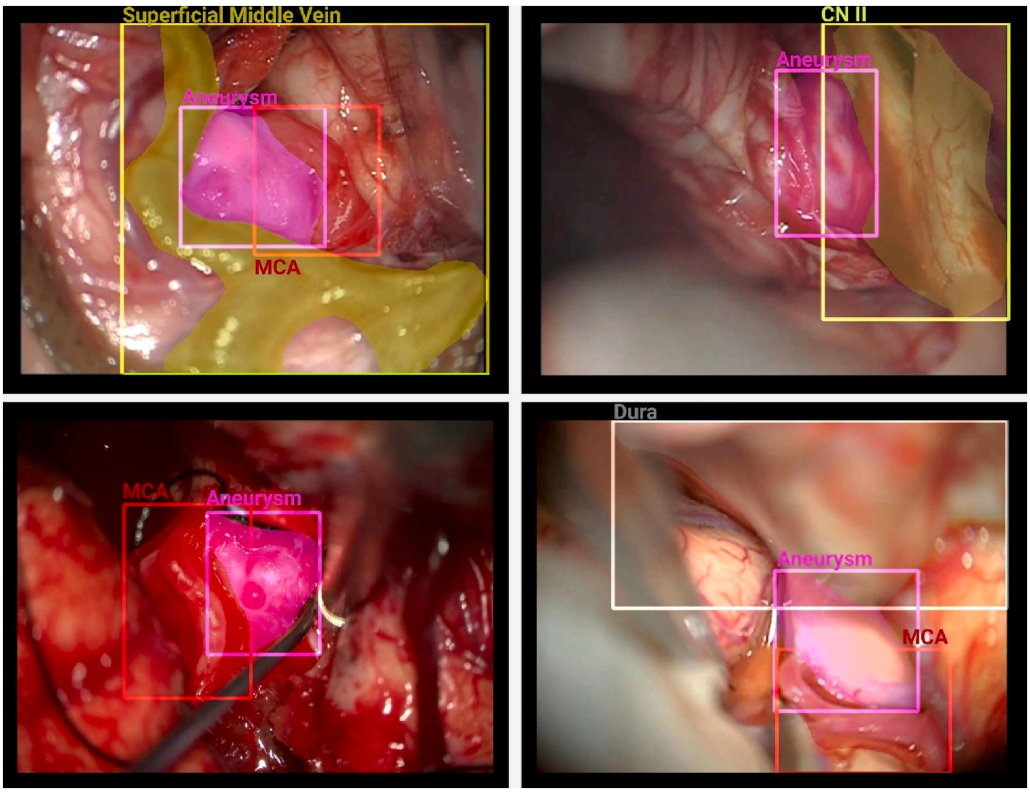

Fig. 11. Segmentation examples for key anatomical structures useful in the context of neurosurgical navigation. The visualizations highlight the model's ability to delineate critical regions such as arterial branches and aneurysms with high precision, even in complex surgical scenes. MCA: Middle Cerebral Artery; CN II: Second Cranial Nerve.

state-of-the-art architectures: our baseline ViT-DeepLab and the Swin Transformer [36]. For each model, we compared: (i) baseline performance using only fully supervised training, (ii) conventional data augmentation with 100% additional synthetic samples, (iii) intensive data augmentation with 200% additional synthetic samples, and (iv) our complete semi-supervised framework with SSU module. The augmentation experiments employed standard techniques including random rotations, brightness/contrast adjustments, and elastic deformations.

Table 6 presents the comparative performance across both architectures with different training approaches, including UniMatch [37], a state-of-the-art semi-supervised learning method. For the Swin Transformer, compared to the fully supervised baseline (60.1% mDice), the integration of our SSU module yielded a relative improvement of + 4.9%, reaching a mean Dice of 63.1%. This improvement surpasses both UniMatch (61.2% mDice) and conventional data augmentation, which provided only marginal benefits with a slight gain at 100% synthetic data (61.3% mDice) and a performance degradation to 59.6% mDice when augmentation was increased to 200%. This pattern was similarly observed with our ViT-DeepLab architecture, where SSU-based pseudo-labeling (+11.3% relative improvement) substantially outperformed both moderate (+1.8%) and intensive (+4.4%) data augmentation.

These findings confirm that more augmentation does not necessarily lead to better results, and may even harm performance when applied excessively. The diminishing or negative returns from intensive data augmentation suggest that synthetic variations alone cannot capture the true distribution complexity of neurosurgical scenes, whereas our uncertainty-guided pseudo-labeling effectively leverages the inherent variability present in real unlabeled data.

Importantly, we also verified that the Semantic Spatial Uncertainty (SSU) metric remained reliable when applied to the Swin Transformer. As in the ViT-based setting, SSU preserved a strong inverse correlation with Dice Similarity Coefficient across all classes (average Pearson correlation of −0.82), supporting its validity as a general-purpose uncertainty estimator independent of the underlying architecture.

While Swin outperformed our baseline architecture in terms of absolute segmentation accuracy, this gain comes at a substantial computational cost. The average inference time for the Swin-based model was

Table 6
Performance comparison of different architectures with and without our semi-supervised framework. Results from UniMatch [37], a state-of-the-art semi-supervised learning method, are included as reference.

Architecture	Training method	mDice	mPrecision	mRecall	Inference time
ViT - DeepLab	Baseline (supervised)	0.503	0.818	0.914	3 ms
	+ 100% Data Augmentation	0.512	0.818	0.914	
	+ 200% Data Augmentation	0.525	0.823	0.910	
	+ SSU module	**0.560**	**0.879**	**0.884**	
Swin	Baseline (supervised)	0.601	0.916	0.922	79 ms
	+ 100% Data Augmentation	0.613	0.917	0.922	
	+ 200% Data Augmentation	0.596	0.910	0.917	
	+ SSU module	**0.631**	**0.927**	**0.910**	
UniMatch [37]	Semi-supervised	0.612	0.891	0.942	45 ms

measured at 79 ms per image, compared to 3 ms per image for our ViT + DeepLabV3 + configuration (measured on an NVIDIA RTX 4090 workstation with PyTorch 2.1.2). This makes Swin approximately 25 times slower, potentially hindering its applicability in real-time surgical settings where high frame rates are essential. In contrast, our ViT-based framework offers a favorable trade-off between accuracy and efficiency, making it more suitable for intraoperative deployment where both segmentation quality and real-time performance are critical requirements.

4. Discussion

Neurosurgical settings have seen an increase in the implementation of computer vision algorithms since their development. In particular, the application of computer vision algorithms relates to real-time anatomical recognition, based on live image feedback from the surgical device, has the potential to serve as a reliable tool for intraoperative orientation. Our study addresses one of the most significant challenges in neurosurgical image analysis: how to develop robust segmentation models with limited annotated data while maintaining clinical reliability. In fact, neurosurgical navigation represents a particularly demanding domain where even small segmentation errors can have significant clinical implications [38].

The experimental results show that our SSU-based framework improves consistently across architectures, implying that reliable estimation methods, rather than architectural innovations, are critical for leveraging unlabeled data. This discovery has significant implications: as new architectures emerge, methods that function independently of specific architectural choices become increasingly valuable. Our findings with both the ViT-DeepLab and Swin implementations support this viewpoint; while Swin achieves higher absolute performance, both architectures improve significantly when enhanced with our framework. This suggests that our method addresses fundamental challenges in neurosurgical data utilization that cannot be solved through architectural modifications alone. Notably, the framework's performance surpasses that of intensive data augmentation (even at 200% additional synthetic samples), confirming that properly filtered real unlabeled data provides more valuable training information than synthetic variations.

The substantial performance improvements observed for rare but clinically critical structures highlight a key strength of our approach—the ability to counteract dataset imbalance through targeted pseudo-labeling. The 40% relative DSC increase for tumor segmentation is particularly meaningful given the critical importance of precise tumor boundary delineation during resection procedures. Similar gains in instrument segmentation (dissectors +42%, clips +19%) suggest that our framework effectively captures the distinctive visual characteristics of these tools despite their infrequent appearance.

An interesting finding was the observed trade-off between precision and recall. The shift toward higher precision after pseudo-labeling reflects the model's evolution toward more conservative predictions, where it only identifies structures when highly confident. The precision improvement demonstrates that our uncertainty-guided approach effectively prevents the common problem of error propagation in pseudo-labeling, where incorrect labels become reinforced through subsequent training iterations [39,40].

The sustained performance improvement across multiple pseudo-labeling iterations validates our adaptive threshold strategy. By recalibrating uncertainty thresholds at each iteration, the model continuously refines its understanding of prediction reliability, avoiding the typical plateau or performance degradation seen in standard pseudo-labeling approaches [41,42]. The diminishing percentage of additional pseudo-labels incorporated in later iterations (2.1% → 1.9% → 1.4%) suggests the model becomes increasingly discriminative, selecting only the most informative examples from the unlabeled pool.

Traditional UQ techniques often fall short in capturing the complexity of multi-class surgical scenes. This limitation is particularly evident in neurosurgical navigation, where some classes are relatively easy to segment (e.g., major blood vessels, lobes) while others present significant challenges due to their rarity or subtle appearance (e.g., tumor boundaries, small vessel branches, cranial nerves). The comparison between SSU and NE revealed that traditional uncertainty metrics perform poorly for structures with high perimeter-to-area ratios, which are common in neurosurgical imagery. SSU shows better correlation with segmentation quality for vessel-like structures, demonstrating improved robustness to spatial characteristics that would otherwise lead to artificially high uncertainty estimates. The proposed SSU metric offers several advantages over traditional uncertainty measures:

- It provides class-specific uncertainty assessment for each anatomical structure or instrument—critical in multi-class neurosurgical scenes where certainty may vary significantly across structures
- It captures spatial uncertainty even when class probabilities are high, identifying regions where boundary positions are uncertain despite confident classification
- It enables interpretable visualization of uncertainty regions, highlighting areas where the model's predictions fluctuate between forward passes

The choice between architectures becomes particularly relevant when considering deployment scenarios. While Swin Transformer achieves higher accuracy (mean DSC improvement of +4.2% compared to ViT-DeepLab), its computational demands vary significantly with hardware configurations. Our benchmarks on professional-grade GPUs (RTX 4090) show Swin's inference time at 75 ms, which increases to 121 ms on consumer hardware (RTX 4060). In contrast, ViT-DeepLab maintains consistent efficiency across hardware configurations (3 ms on professional GPUs, 6 ms on consumer hardware). This performance scaling has important implications for clinical deployment: while Swin's superior accuracy makes it ideal for offline processing tasks such as surgical video analysis or educational content creation, ViT-DeepLab's consistent low latency makes it more suitable for real-time applications where maintaining surgical workflow is crucial.

Despite these promising results, our approach shows limitations in addressing out-of-distribution examples. The pseudo-labeling selection process inherently excludes surgical scenes that differ significantly from the training distribution, suggesting that model improvement through semi-supervised learning has natural boundaries without additional diversity in the labeled dataset. Additionally, being trained on single-center data, the model's generalizability to different surgical settings and institutional protocols remains to be validated. Future work should explore combining our approach with active learning strategies to identify the most informative outlier cases for targeted annotation, and extend validation to multi-center settings. Additionally, our current implementation analyzes each frame independently and does not incorporate temporal consistency across video sequences. This limitation stems from our data collection strategy, which prioritized annotating diverse individual frames over continuous sequences to maximize the variety of captured surgical scenarios. Future development could leverage inter-frame relationships to further enhance segmentation coherence and robustness, potentially enabling more accurate tracking of structures throughout surgical procedures.

To our knowledge, this work represents the first uncertainty-aware semi-supervised segmentation framework optimized for neurosurgical navigation. By reframing the annotation scarcity problem as an uncertainty quantification challenge, our approach offers a practical pathway to enhance segmentation performance while maintaining the reliability necessary for clinical deployment.

5. Conclusion

We presented a novel uncertainty-aware semi-supervised framework for neurosurgical video segmentation that effectively leverages

unlabeled data to enhance performance in this data-limited domain. Our approach introduces Semantic Spatial Uncertainty (SSU), a new metric that correlates strongly with segmentation quality and enables reliable pseudo-label selection without ground truth. Through adaptive calibration and class-specific uncertainty thresholds, our method achieves substantial improvements for clinically critical structures (+40% relative DSC for tumors, + 14% for aneurysms), demonstrating particular value for underrepresented classes. Future work should address current limitations by incorporating temporal consistency across video frames and exploring active learning to identify informative outlier cases for annotation. This approach represents a promising direction for medical image segmentation tasks where annotated data is scarce but reliable predictions are essential for clinical applications.

CRediT authorship contribution statement

Francesco Nitti: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Silvia Seoni:** Writing – review & editing, Software, Methodology, Formal analysis. **Alberto Morello:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Lorenzo Dolci:** Writing – review & editing, Investigation. **Amedeo Piazza:** Writing – review & editing, Investigation. **Vincenzo Esposito:** Writing – review & editing, Investigation. **Luigi Rosito:** Writing – review & editing, Investigation. **Diego Garbossa:** Writing – review & editing, Investigation. **Fabio Cofano:** Writing – review & editing, Investigation, Data curation. **Abdulkadir Sengur:** Writing – review & editing. **Massimo Salvi:** Writing – original draft, Validation, Supervision, Resources, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Open access publishing facilitated by Politecnico di Torino, as part of the Elsevier - CRUI-CARE agreement.

Data availability

We have created an open source repository with the entire implementation of our codes (the link is provided in the manuscript)

References

[1] J. Bellomo, A. Spinello, A. Morello, T. Schubert, J. Fierstra, A. Piazza, C. Serra, Human brain vasculature, in: Encyclopedia of the Human Brain, Elsevier, 2025, pp. 62–87, https://doi.org/10.1016/B978-0-12-820480-1.00191-1.
[2] I.J. Gerard, M. Kersten-Oertel, J.A. Hall, D. Sirhan, D.L. Collins, Brain shift in neuronavigation of brain tumors: an updated review of intra-operative ultrasound applications, Front. Oncol. 10 (2021) 618837, https://doi.org/10.3389/fonc.2020.618837.
[3] C. Schulz, S. Waldeck, U.M. Mauer, Intraoperative image guidance in neurosurgery: development, current indications, and future trends, Radiol. Res. Pract. 2012 (2012) 1–9, https://doi.org/10.1155/2012/197364.
[4] T. Mitsui, M. Fujii, M. Tsuzaka, Y. Hayashi, Y. Asahina, T. Wakabayashi, Skin shift and its effect on navigation accuracy in image-guided neurosurgery, Radio. Phys. Technol. 4 (2011) 37–42, https://doi.org/10.1007/s12194-010-0103-0.
[5] V.E. Staartjes, G. Sarwin, A. Carretta, M. Zoli, D. Mazzatenta, L. Regli, E. Konukoglu, C. Serra, AENEAS project: live image-based navigation and roadmap generation in endoscopic neurosurgery using machine vision, Oper. Neurosurg. (2025), https://doi.org/10.1227/ons.0000000000001583.
[6] S. Drouin, A. Kochanowska, M. Kersten-Oertel, I.J. Gerard, R. Zelmann, D. De Nigris, S. Bériault, T. Arbel, D. Sirhan, A.F. Sadikot, J.A. Hall, D.S. Sinclair, K. Petrecca, R.F. DelMaestro, D.L. Collins, IBIS: an OR ready open-source platform for image-guided neurosurgery, Int J. CARS 12 (2017) 363–378, https://doi.org/10.1007/s11548-016-1478-0.
[7] G. Sarwin, A. Carretta, V. Staartjes, M. Zoli, D. Mazzatenta, L. Regli, C. Serra, E. Konukoglu, Live image-based neurosurgical guidance and roadmap generation using unsupervised embedding, in: A. Frangi, M. De Bruijne, D. Wassermann, N. Navab (Eds.), Information Processing in Medical Imaging, Springer Nature Switzerland, Cham, 2023, pp. 107–118, https://doi.org/10.1007/978-3-031-34048-2_9.
[8] M. Islam, D.A. Atputharuban, R. Ramesh, H. Ren, Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning, IEEE Robot. Autom. Lett. 4 (2019) 2188–2195, https://doi.org/10.1109/LRA.2019.2900854.
[9] R. Azad, E.K. Aghdam, A. Rauland, Y. Jia, A.H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J.P. Cohen, E. Adeli, D. Merhof, Medical image segmentation review: the success of U-net, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2024) 10076–10095, https://doi.org/10.1109/TPAMI.2024.3435571.
[10] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang, X. Liu, J. Chen, H. Zhou, I. Ben Ayed, H. Zheng, Annotation-efficient deep learning for automatic medical image segmentation, Nat. Commun. 12 (2021) 5915, https://doi.org/10.1038/s41467-021-26216-9.
[11] Z. Solatidehkordi, I. Zualkernan, Survey on recent trends in medical image classification using semi-supervised learning, Appl. Sci. 12 (2022) 12094, https://doi.org/10.3390/app122312094.
[12] Y. Wu, Z. Ge, D. Zhang, M. Xu, L. Zhang, Y. Xia, J. Cai, Mutual consistency learning for semi-supervised medical image segmentation, Med. Image Anal. 81 (2022) 102530, https://doi.org/10.1016/j.media.2022.102530.
[13] J. Yang, H. Li, H. Wang, M. Han, 3D medical image segmentation based on semi-supervised learning using deep co-training, Appl. Soft Comput. 159 (2024) 111641, https://doi.org/10.1016/j.asoc.2024.111641.
[14] K. Han, V.S. Sheng, Y. Song, Y. Liu, C. Qiu, S. Ma, Z. Liu, Deep semi-supervised learning for medical image segmentation: a review, Expert Syst. Appl. 245 (2024) 123052, https://doi.org/10.1016/j.eswa.2023.123052.
[15] Y. Fan, A. Kukleva, D. Dai, B. Schiele, Revisiting consistency regularization for semi-supervised learning, Int J. Comput. Vis. 131 (2023) 626–643, https://doi.org/10.1007/s11263-022-01723-4.
[16] Y. Jo, H. Kahng, S.B. Kim, Deep semi-supervised regression via pseudo-label filtering and calibration, Appl. Soft Comput. 161 (2024) 111670, https://doi.org/10.1016/j.asoc.2024.111670.
[17] J. Zhu, C. Huang, H. Xi, H. Cui, CCA: contrastive cluster assignment for supervised and semi-supervised medical image segmentation, Neural Netw. 188 (2025) 107415, https://doi.org/10.1016/j.neunet.2025.107415.
[18] Q. Zeng, Y. Xie, Z. Lu, M. Lu, Y. Wu, Y. Xia, Segment together: a versatile paradigm for semi-supervised medical image segmentation, IEEE Trans. Med. Imaging 44 (2025) 2948–2959, https://doi.org/10.1109/TMI.2025.3556310.
[19] F. Tang, Z. Xu, M. Hu, W. Li, P. Xia, Y. Zhong, H. Wu, J. Su, Z. Ge, Neighbor does matter: density-aware contrastive learning for medical semi-supervised segmentation, AAAI 39 (2025) 7220–7228, https://doi.org/10.1609/aaai.v39i7.32776.
[20] L. Yang, Z. Zhao, H. Zhao, UniMatch V2: pushing the limit of semi-supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 47 (2025) 3031–3048, https://doi.org/10.1109/TPAMI.2025.3528453.
[21] Y. Chen, Y. Liu, M. Lu, L. Fu, F. Yang, Multi-consistency for semi-supervised medical image segmentation via diffusion models, Pattern Recognit. 161 (2025) 111216, https://doi.org/10.1016/j.patcog.2024.111216.
[22] K. Wang, C. Zhang, Y. Geng, H. Ma, Evidential pseudo-label ensemble for semi-supervised classification, Pattern Recognit. Lett. 177 (2024) 135–141, https://doi.org/10.1016/j.patrec.2023.11.027.
[23] M. Salvi, S. Seoni, A. Campagner, A. Gertych, U.R. Acharya, F. Molinari, F. Cabitza, Explainability and uncertainty: two sides of the same coin for enhancing the interpretability of deep learning models in healthcare, Int. J. Med. Inform. 197 (2025) 105846, https://doi.org/10.1016/j.ijmedinf.2025.105846.
[24] W. Huang, L. Zhang, Z. Wang, L. Wang, Exploring inherent consistency for semi-supervised anatomical structure segmentation in medical imaging, IEEE Trans. Med. Imaging 43 (2024) 3731–3741, https://doi.org/10.1109/TMI.2024.3400840.
[25] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30 (2017).
[26] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: PMLR, 2016, pp. 1050–1059.
[27] K. Wu, H. Peng, M. Chen, J. Fu, H. Chao, Rethinking and improving relative position encoding for vision transformer, in: 2021: pp. 10033–10041.
[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder, Atrous Separable Convolution Semant. Image Segm. (2018), https://doi.org/10.48550/ARXIV.1802.02611.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
[30] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. Ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (1987) 355–368, https://doi.org/10.1016/S0734-189X(87)80186-X.
[31] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X.X. Zhu, A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (2023) 1513–1589, https://doi.org/10.1007/s10462-023-10562-9.
[32] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297, https://doi.org/10.1016/j.inffus.2021.05.008.

[33] M. Salvi, A. Mogetta, U. Raghavendra, A. Gudigar, U.R. Acharya, F. Molinari, A dynamic uncertainty-aware ensemble model: application to lung cancer segmentation in digital pathology, Appl. Soft Comput. 165 (2024) 112081, https://doi.org/10.1016/j.asoc.2024.112081.
[34] F.C. Maruccio, W. Eppinga, M.-H. Laves, R.F. Navarro, M. Salvi, F. Molinari, P. Papaconstadopoulos, Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation, Phys. Med. Biol. 69 (2024) 035007, https://doi.org/10.1088/1361-6560/ad1a26.
[35] M. Yan, S.C. Hui, N. Li, DML-PL: deep metric learning based pseudo-labeling framework for class imbalanced semi-supervised learning, Inf. Sci. 626 (2023) 641–657, https://doi.org/10.1016/j.ins.2023.01.074.
[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021: pp. 10012–10022.
[37] L. Yang, L. Qi, L. Feng, W. Zhang, Y. Shi, Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation, (2022). ⟨https://doi.org/10.48550/ARXIV.2208.09910⟩.
[38] P.W.A. Willems, J.W.B. Van Der Sprenkel, C.A.F. Tulleken, M.A. Viergever, M.J. B. Taphoorn, Neuronavigation and surgery of intracerebral tumours, J. Neurol. 253 (2006) 1123–1136, https://doi.org/10.1007/s00415-006-0158-3.
[39] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, Y. Gao, Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation, IEEE Trans. Med. Imaging 41 (2022) 608–620, https://doi.org/10.1109/TMI.2021.3117888.
[40] L. Lu, M. Yin, L. Fu, F. Yang, Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation, Biomed. Signal Process. Control 79 (2023) 104203, https://doi.org/10.1016/j.bspc.2022.104203.
[41] L.-Z. Guo, L.-H. Jia, J.-J. Shao, Y.-F. Li, Robust semi-supervised learning in open environments, Front. Comput. Sci. 19 (2025) 198345, https://doi.org/10.1007/s11704-024-40646-w.
[42] L.-Z. Guo, Y.-F. Li, Class-imbalanced semi-supervised learning with adaptive thresholding, in: PMLR, 2022, pp. 8082–8094.