Multi-Task Semi-Supervised Learning for Intracranial Hemorrhage Analysis in Brain CT

Phuong-Anh Nguyen[1], Thi-Lan Le[1], and Thanh-Hai Tran[1]

[1]School of Electrical and Electronic Engineering
Hanoi University of Science and Technology
anhntp8701@gmail.com, lan.lethi1@hust.edu.vn, thanh-hai.tran@mica.edu.vn

Abstract. Intracranial hemorrhage (ICH) analysis requires both accurate lesion localization and reliable subtype recognition. While deep learning has advanced ICH analysis, existing methods address segmentation and classification as independent tasks, overlooking the strong anatomical correlation between hemorrhage extent and subtype identity, and are further constrained by the scarcity of pixel-level annotations. We propose a two-phase framework for joint binary segmentation (hemorrhage vs. background) and multi-label subtype classification from 2D CT slices. In the supervised phase, a shared Wide-ResNet50-2 encoder is coupled with a UNet++ segmentation decoder and a lesion-guided classification head, where predicted hemorrhage masks serve as soft spatial priors for subtype recognition. In the semi-supervised phase, a warm-start Mean Teacher strategy leverages unlabeled CT slices through consistency regularization with confidence filtering applied only to the segmentation branch, while classification benefits indirectly through the shared encoder. Experiments on the Brain Hemorrhage Segmentation Dataset (BHSD) show that multi-task learning improves Dice by 11.48% absolute and macro-F1 by 5.45% absolute over single-task baselines. The semi-supervised stage further improves performance to 77.87% Dice and 65.50% macro-F1, demonstrating that segmentation-guided multi-task learning and unlabeled-data regularization are complementary for CT-based ICH analysis.

Keywords: Intracranial hemorrhage · Multi-task learning · Semi-supervised learning · Hemorrhage segmentation

1 Introduction

Intracranial hemorrhage (ICH) is one of the most severe subtypes of stroke, accounting for approximately 10–15% of all stroke cases worldwide, yet carrying a disproportionately high mortality rate of 40–50% within the first month [1]. Non-contrast computed tomography (CT) remains the gold standard imaging modality for ICH assessment due to its wide availability and rapid acquisition. However, manual delineation of hemorrhage regions by radiologists is labor-intensive and increasingly impractical given the growing volume of emergency imaging studies.

Deep learning has shown remarkable promise for automating ICH analysis. Despite these advances, two fundamental challenges remain underaddressed. First, most methods treat *segmentation* and *classification* as isolated tasks. This ignores the inherent complementarity between pixel-level hemorrhage localization and image-level subtype recognition: the spatial extent and density of hemorrhagic regions are strong indicators of hemorrhage subtype, while subtype-specific semantic knowledge can guide more precise boundary delineation. Training both tasks independently also misses opportunities for shared anatomical representation learning. Second, expert-delineated hemorrhage masks are expensive to obtain, resulting in small labeled datasets. In contrast, large quantities of unannotated CT scans are routinely acquired and archived, motivating the use of *semi-supervised learning* (SSL) to exploit this data.

To address both challenges simultaneously, we propose a unified framework integrating multi-task learning and semi-supervised learning for ICH analysis. Our approach is structured in two sequential phases as illustrated in Fig. 1. Beyond these two core challenges, a further practical question arises in CT-based deep learning that is often overlooked: *which Hounsfield Unit (HU) window is most informative for hemorrhage analysis?* CT scans encode tissue density across a wide HU range, and clinical reading routinely employs multiple display windows — each highlighting different anatomical structures: the brain window emphasizes parenchymal contrast, the subdural window enhances extra-axial collections, and the bone window visualizes skull integrity. Yet most deep learning methods for ICH adopt a single default window (typically brain) without systematic justification, potentially discarding diagnostically relevant signal. We therefore include a controlled study of windowing as input representation among our contributions.

Our main contributions are as follows:

- **Lesion-guided multi-task architecture.** We design a shared Wide-ResNet50-2 encoder coupled with a UNet++ segmentation decoder and a novel lesion-guided classification head that explicitly uses the predicted hemorrhage mask as a spatial attention prior, enabling simultaneous segmentation and five-subtype multi-label classification (Intraventricular, Intraparenchymal, Subarachnoid, Epidural, and Subdural).
- **CT windowing ablation.** We conduct a systematic study of windowing configurations as input representations and empirically identify the optimal two-channel dual-window strategy that outperforms both single-window and three-window alternatives.
- **Semi-supervised segmentation regularization.** We extend the multi-task framework with a warm-start Mean Teacher SSL pipeline that leverages unlabeled CT slices via confidence-aware consistency regularization applied exclusively to the segmentation branch, propagating indirect gains to classification through the shared encoder.

Experiments on BHSD[2] demonstrate that multi-task learning yields consistent gains over single-task baselines, and the SSL stage further improves these

performances. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 presents our proposed method; Section 4 describes experimental settings and results; and Section 5 concludes the paper.

2 Related Work

2.1 ICH Segmentation

Early automated approaches to ICH analysis relied on hand-crafted features with classical classifiers [14]. The advent of deep learning brought substantial improvements: Chilamkurthy et al. [15] demonstrated that CNNs could detect critical head CT findings at radiologist-level accuracy, though their method operates at the image level without producing pixel-wise segmentation masks. For segmentation, encoder–decoder architectures based on U-Net [3] quickly became the dominant paradigm, leveraging skip connections to preserve spatial detail essential for delineating hemorrhage boundaries. Subsequent 2D methods such as TransUNet [17] and the lightweight UNeXt [16] further improved performance on binary ICH segmentation. Wang et al. [18] recently proposed DE-UNeXt, a dual-encoder architecture that feeds both the original image and its inverted image into parallel encoders to capture complementary features, achieving strong binary segmentation performance on BHSD.

Despite these advances, most existing methods treat ICH as a binary segmentation problem, ignoring clinically important subtype distinctions. Wu et al. [2] introduced the BHSD dataset and benchmarked five 3D models (UNETR, Swin UNETR, CoTr, nnFormer, nnUNet) on the ICH segmentation task. However, these 3D approaches require volumetric sub-volume cropping, which substantially reduces the effective number of training samples and demands high GPU memory, limiting practical deployment. Our work addresses this gap by operating on 2D axial slices and jointly learning binary hemorrhage segmentation with five-subtype multi-label classification.

2.2 Multi-Task Learning for Medical Imaging

Multi-task learning (MTL) aims to improve generalization by jointly optimizing related tasks through shared representations [6]. In medical imaging, Zhou et al. [8] demonstrated consistent gains from joint segmentation and classification for breast ultrasound tumor analysis, showing that pixel-level localization and image-level diagnosis mutually reinforce each other. For brain imaging, Mlynarski et al. [9] combined tumor segmentation with survival prediction, achieving improvements in both tasks over single-task baselines. A critical challenge in MTL is balancing gradient contributions across tasks to prevent one task from dominating training. Adaptive strategies such as Uncertainty Weighting [7] have been shown to yield more stable convergence than fixed-weight approaches by learning per-task loss weights automatically.

Despite the success of MTL in related medical imaging domains, its application to ICH analysis remains largely unexplored. Existing ICH methods treat

segmentation and subtype classification as independent pipelines, missing the opportunity to leverage the strong anatomical correlation between hemorrhage extent and subtype identity. Our work exploits this complementarity through a novel lesion-guided classification head that uses predicted segmentation masks as spatial attention priors.

2.3 Semi-Supervised Learning for Medical Segmentation

Semi-supervised learning (SSL) addresses the annotation bottleneck in medical imaging by exploiting large quantities of unannotated data alongside a small labeled set. Mean Teacher [11] maintains an exponential moving average (EMA) of student weights as a teacher that generates reliable pseudo-labels for unlabeled inputs. FixMatch [12] extended this idea by applying a confidence threshold to filter uncertain pseudo-labels, ensuring only high-quality targets contribute to training. In medical image segmentation, UA-MT [13] adapted Mean Teacher with uncertainty-aware pseudo-label selection for 3D cardiac segmentation.

For ICH specifically, Wu et al. [2] benchmarked four SSL methods on BHSD — Mean Teacher, Cross Pseudo Supervision (CPS), Entropy Minimization, and Interpolation Consistency Training — finding that CPS achieved the best binary segmentation Dice of 49.5% among semi-supervised approaches. However, these methods operate in a single-task segmentation setting and do not exploit the complementarity between segmentation and classification. Furthermore, cold-start SSL on medical data risks unreliable early pseudo-labels that can destabilize training. Our work addresses both limitations by applying consistency regularization selectively to the segmentation branch of a pre-trained multi-task model, using warm-start initialization to ensure pseudo-label quality from the first SSL epoch.

3 Proposed Method

Fig. 1 provides an overview of our proposed framework, which consists of two sequential training phases: (i) a *supervised multi-task learning* phase on labeled CT slices, and (ii) a *semi-supervised* phase that extends the trained model using unlabeled data via Mean Teacher consistency regularization. Given a 2D CT slice x, the segmentation branch predicts a binary hemorrhage mask $\hat{m} \in [0,1]^{1\times H\times W}$, where foreground denotes any ICH subtype. The classification branch predicts a multi-label subtype vector $\hat{\mathbf{y}} \in [0,1]^5$, since multiple hemorrhage subtypes may co-occur in the same slice.

3.1 CT Input Preprocessing

CT scans encode voxel intensities in Hounsfield Units (HU), which exceed the dynamic range of standard 8-bit representations. *Windowing* is a clinical technique that maps a selected HU range to the full display range, highlighting specific tissue contrasts. Different window configurations emphasize different hemorrhage

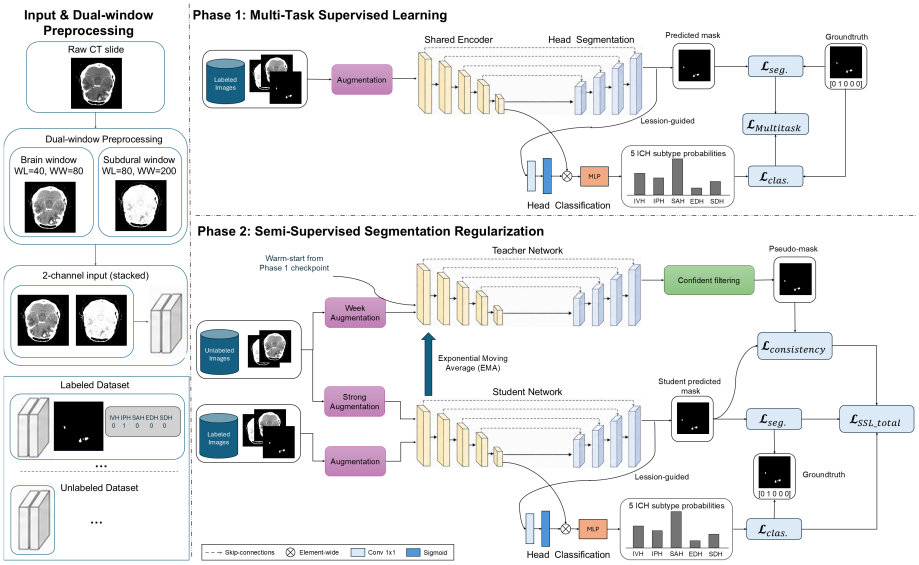

Fig. 1: Overview of the proposed two-phase framework. **Phase 1**: A shared Wide-ResNet50-2 encoder jointly trains a UNet++ segmentation decoder and a lesion-guided classification head. **Phase 2**: Both student and teacher (EMA of student) are initialized from the Phase 1 checkpoint. Consistency regularization is applied exclusively to the segmentation branch using unlabeled CT slices.

types: the brain window reveals parenchymal abnormalities, while the subdural window enhances extra-axial collections.

We adopt a *dual-window* strategy, selecting the two most informative windows as separate input channels. For each CT slice, we apply:

$$x^{(k)} = \mathrm{clip}\left(\frac{I - (c_k - w_k/2)}{w_k},\ 0,\ 1\right), \quad k = 1, 2 \tag{1}$$

where I is the raw HU map, and (c_k, w_k) are the center and width of each window: **brain** (c=40, w=80) and **subdural** (c=80, w=200). The two normalized maps are concatenated to form a two-channel input $x = [x^{(1)}, x^{(2)}] \in \mathbb{R}^{2\times H\times W}$. The choice of this dual-window configuration is motivated by our ablation study (Section 4.2), which empirically demonstrates its superiority over single-window and three-window alternatives. Since the Wide-ResNet50-2 encoder is pretrained on three-channel ImageNet inputs, we adapt the first convolutional layer to our two-channel CT input by retaining the pretrained weights of the first two input channels and discarding the third. All subsequent encoder layers are initialized with the original ImageNet-pretrained weights.

3.2 Phase 1: Multi-Task Supervised Learning

Motivation. Existing ICH methods typically address segmentation and classification independently. This misses two important synergies: (i) both tasks share anatomical representations that can be jointly learned, and (ii) hemorrhage location information directly informs subtype recognition. We propose a unified multi-task framework that exploits these synergies through a shared encoder and a novel lesion-guided classification design.

Shared Encoder. We adopt Wide-ResNet50-2 [5] pretrained on ImageNet as the shared backbone. Given an input slice, the encoder produces a hierarchy of multi-scale feature maps $\mathcal{F}$.

These features serve as the basis for both segmentation and classification, enabling implicit knowledge transfer between the two tasks.

Segmentation Branch. The feature hierarchy $\mathcal{F}$ is passed to a UNet++ decoder [4], which replaces the sparse skip connections of standard U-Net with dense nested convolutional blocks, facilitating multi-scale feature aggregation and precise boundary delineation. The decoder outputs a pixel-wise probability map:

$$\hat{m} = \text{SegHead}(\text{UNet++Decoder}(\mathcal{F})) \in [0,1]^{1\times H\times W} \tag{2}$$

Lesion-Guided Classification Head. A common multi-task design in medical imaging uses image-level classification predictions to guide pixel-level segmentation — the coarse output of the simpler task informs the finer task. We propose the inverse coupling: *pixel-level segmentation predictions guide image-level classification*. This direction is motivated by an information asymmetry — a pixel-wise mask carries substantially richer spatial evidence (location, extent, shape, boundary) than an image-level label, and is therefore more informative as a prior. To exploit the richer spatial evidence in segmentation masks, we couple the tasks in the segmentation-to-classification direction.

Concretely, the predicted hemorrhage probability map $\hat{m}$ is transformed into a soft spatial attention map $\mathcal{A}$, which highlights regions likely to contain hemorrhage:

$$\mathcal{A} = \sigma(\text{Conv}_{1\times1}(\hat{m})) \in [0,1]^{1\times H\times W} \tag{3}$$

This attention map reweights the high-level encoder feature f before global average pooling:

$$z = \text{GAP}(f \odot \mathcal{A}) \in \mathbb{R}^{2048} \tag{4}$$

The Hadamard product $f \odot \mathcal{A}$ suppresses background activations (where $\mathcal{A} \approx 0$) and amplifies lesion-aligned activations (where $\mathcal{A} \approx 1$). The resulting lesion-aware representation z is passed through a multi-layer perceptron (MLP) with batch normalization and dropout to produce multiple ICH subtypes predictions:

$$\hat{\mathbf{y}} = \sigma(MLP(z)) \in [0,1]^C, \quad C = 5. \tag{5}$$

Crucially, gradients from the classification loss flow back through $\mathcal{A}$ into $\hat{m}$, so segmentation also benefits from classification supervision. The resulting coupling is therefore *bidirectional* rather than a one-way prior: the two tasks co-adapt during training. This design encourages the classifier to focus on anatomically relevant regions, reducing background noise and strengthening the link between localization and diagnosis.

Multi-Task Loss. For segmentation, we use Dice Loss that directly optimizes overlap:

$$\mathcal{L}_{\text{seg}} = 1 - \frac{2\sum_i \hat{p}_i\, y_i + \epsilon}{\sum_i \hat{p}_i + \sum_i y_i + \epsilon} \tag{6}$$

For classification, we apply a class-balanced focal loss [10] to address label imbalance across five ICH subtypes:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{BC}\sum_{b=1}^{B}\sum_{c=1}^{C} \alpha_c\,(1-p_t^{bc})^{\gamma}\log(p_t^{bc}) \tag{7}$$

with per-class weights $\alpha_c = N_c^{\text{neg}}/N_c^{\text{pos}}$ and γ=2.0. The total supervised loss uses uncertainty-based weighting [7] to balance gradient contributions:

$$\mathcal{L}_{\text{sup}} = \frac{1}{2e^{\hat{s}_1}}\mathcal{L}_{\text{seg}} + \hat{s}_1 + \frac{1}{2e^{\hat{s}_2}}\mathcal{L}_{\text{cls}} + \hat{s}_2 \tag{8}$$

where $\hat{s}_1, \hat{s}_2$ are learnable log-variance parameters.

3.3 Phase 2: Semi-Supervised Segmentation Regularization

Warm-Start Mean Teacher. We first train the Phase 1 multi-task model to convergence, obtaining checkpoint θ^*_{SL}. Both student f_{θ_s} and teacher f_{θ_t} are initialized from θ^*_{SL} — a *warm-start* strategy that ensures reliable teacher pseudo-labels from the first SSL training step. The teacher is updated as an exponential moving average of the student:

$$\theta_t^{(k)} = \alpha\,\theta_t^{(k-1)} + (1-\alpha)\,\theta_s^{(k)}, \tag{9}$$

We set the EMA decay rate $\alpha = 0.999$, following standard practice in the Mean Teacher literature [11]. This high value ensures that the teacher evolves slowly relative to the student, producing stable pseudo-labels that average out noise in the student's stochastic updates.

Confidence-Aware Consistency Loss. For each unlabeled image u_j, the teacher generates a soft pseudo-mask under weak augmentation. We retain only confident pixels via a FixMatch-style threshold [12]:

$$\mathbf{c}_{ij} = \mathbb{1}[\tilde{m}_{ij} > \tau \;\text{ or }\; \tilde{m}_{ij} < 1-\tau], \quad \tau = 0.8 \tag{10}$$

Table 1: BHSD hemorrhage slice distribution after 2D extraction.

Subtype	**Train slices**	**Test slices**
EDH (Epidural)	63	118
SDH (Subdural)	344	421
IVH (Intraventricular)	355	358
IPH (Intraparenchymal)	441	447
SAH (Subarachnoid)	488	488
Total (positive slices)	1,114	1,254
Background-only	557	1,973
Unlabeled	18,969	-

The consistency loss is a masked BCE between student predictions and binarized teacher pseudo-labels:

$$\mathcal{L}_{\text{cons}} = \frac{\sum_{i,j} \mathbf{c}_{ij} \cdot \text{BCE}\big(f_{\theta_s}(u_j^{\text{strong}})_i,\, \mathbb{1}[\tilde{m}_{ji} > 0.5]\big)}{\sum_{i,j} \mathbf{c}_{ij} + \epsilon} \tag{11}$$

Total SSL Loss and Ramp-Up. A linear ramp-up schedule gradually increases the unsupervised term, preventing noisy early pseudo-labels from disrupting the supervised objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda(t) \cdot \mathcal{L}_{\text{cons}}, \quad \lambda(t) = \lambda_{\max} \cdot \min\left(1,\, \frac{t}{T_{\text{ramp}}}\right) \tag{12}$$

where T_{ramp}=5 epochs and $\lambda_{\max}$=1.0.

4 Experiments and Results

4.1 Dataset and Implementation

Dataset. We conduct experiments on the Brain Hemorrhage Segmentation Dataset (BHSD) [2], a publicly available 3D CT dataset comprising 192 volumes with full pixel-level annotations across five ICH subtypes, including Epidural (EDH), Subdural (SDH), Intraventricular (IVH), Intraparenchymal (IPH), and Subarachnoid (SAH), and 562 additional volumes with confirmed hemorrhage presence but without slice-level annotations, used as the unlabeled pool for SSL. We follow the official train/test split and decompose 3D volumes into 2D axial slices of size 512×512. For supervised training, we use all hemorrhage-positive slices and randomly sample background-only slices from the labeled training volumes. For evaluation, we report metrics on the full test set, which includes both foreground and background-only slices, to reflect realistic deployment conditions. Table 1 summarizes the resulting per-subtype distribution.

Implementation Details. All experiments are implemented in PyTorch on a single NVIDIA A100 GPU. The Wide-ResNet50-2 [5] backbone is initialized from ImageNet pretrained weights. Phase 1 uses AdamW with learning rate 5×10^{-4} and CosineAnnealing scheduler for 100 epochs. Phase 2 uses Adam at 1×10^{-4} for 150 epochs with early stopping. Effective batch size of 32 is achieved via gradient accumulation.

Evaluation Metrics We use Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) for segmentation task and F1 score for classification task.

4.2 Ablation Study: CT Windowing Configuration

Table 2 reports segmentation and classification performance for different windowing configurations as input to the Phase 1 multi-task model. We evaluate single windows (brain, subdural, bone), the two-channel brain+subdural combination, and the three-channel combination. The dual-window Brain+Subdural combination consistently outperforms all single-window and the three-window configuration. The brain window captures parenchymal hemorrhage contrast, while the subdural window enhances extra-axial collections; together they provide complementary information. Adding the bone window (designed for skull visualization) introduces irrelevant features that slightly degrade performance, suggesting that careful window selection is more important than maximizing channel count.

Table 2: Ablation study on CT windowing configurations. All models use the same multi-task architecture. Window settings: Brain (c=40, w=80), Subdural (c=80, w=200), Bone (c=500, w=2500).

Input Configuration	**DSC↑**	**IoU↑**	**F1$_{\text{mac}}$↑**
Brain only	0.7181	0.6771	0.6083
Subdural only	0.7248	0.6852	0.5727
Bone only	0.6846	0.6296	0.5537
Brain + Subdural	**0.7470**	**0.7081**	**0.6164**
Brain + Subdural + Bone	0.7138	0.6714	0.5689

4.3 Ablation study: Multi-Task vs. Single-Task

Table 3 compares our multi-task framework against task-specific single-task baselines, all using the same backbone and dual-window input. The results validate our core motivation: jointly training segmentation and classification outperforms either task trained in isolation. Multi-task learning yields substantial improvements: 11.48% absolute in Dice and 21.88% absolute in IoU over the segmentation-only baseline, and 5.45% absolute in F1 over the classification-only

Table 3: Comparison of single-task and multi-task learning under fully supervised training on BHSD. All models use Brain+Subdural dual-window input and Wide-ResNet50-2 encoder.

Method	**DSC↑**	**IoU↑**	**F1$_{mac}$↑**
Classification only	—	—	0.5619
Segmentation only	0.6322	0.4893	—
Multi-task (ours)	**0.7470**	**0.7081**	**0.6164**

baseline. These gains stem from the shared encoder learning jointly enriched representations, and from the lesion-guided classification head that leverages segmentation predictions as spatial priors.

4.4 Results: Semi-Supervised Improvement

Table 4 reports the effect of Phase 2 SSL training on top of the Phase 1 multi-task model. The semi-supervised stage consistently improves all metrics: 3.2% absolute in Dice, 3.5% absolute in IoU, and 3.9% absolute in macro-F1. Notably, classification improves despite the consistency loss being applied *only* to the segmentation branch — confirming our design hypothesis that improved segmentation representations propagate through the shared encoder to benefit classification.

Table 4: Effect of semi-supervised learning on top of multi-task supervised training.

Method	**DSC↑**	**IoU↑**	**F1↑**
Multi-task (fully supervised)	0.7470	0.7081	0.6164
Multi-task + SSL (ours)	**0.7787**	**0.7427**	**0.6550**

To further analyze the source of the macro-F1 improvement, Table 5 reports per-subtype classification F1 across the three training settings. Multi-task learning improves classification performance for every subtype over the classification-only baseline, with particularly notable gains for EDH and SDH (+0.073 F1 each). This supports the hypothesis that spatial priors derived from segmentation help the classifier distinguish anatomically localized hemorrhage patterns, especially for extra-axial subtypes where spatial location is highly diagnostic. The semi-supervised stage further improves all subtypes, suggesting that improved segmentation representations are especially beneficial when direct classification supervision is scarce. Nevertheless, EDH remains the most challenging subtype in absolute terms, reaching only 0.292 F1 after SSL. This reflects the severe class imbalance and limited visual diversity of EDH examples, and suggests that rare-subtype classification remains an important direction for future work.

Table 5: Per-subtype classification F1 under classification-only, Phase 1: Supervised multi-task, and Phase 2: Semi-supervised multi-task settings. Train slices indicate the number of labeled foreground slices available for each subtype.

Subtype	Train slices	Classification-only	Phase 1	Phase 2
EDH	63	0.159	0.232	0.292
SDH	344	0.556	0.629	0.649
IVH	355	0.716	0.755	0.794
IPH	441	0.765	0.824	0.854
SAH	488	0.613	0.645	0.685

4.5 Comparison with State-of-the-Art Methods

Table 6: Comparison with state-of-the-art methods on BHSD (binary Dice). †Official BHSD benchmark [2]. ‡Results reported in [18].

Paradigm	**Method**	**Input**	**DSC↑**
Supervised	U-Net [3]‡	2D	0.468
	TransUNet [17]‡	2D	0.528
	UNeXt [16]‡	2D	0.669
	DE-UNeXt [18]‡	2D	0.704
	nnUNet [2]	3D	0.451
Semi-supervised	CPS [2]	3D	0.495
	Multi-task + SSL (ours)	2D	**0.779**
Supervised ablation (ours):			
	Multi-task (Phase 1 only)	2D	0.747

Table 6 compares our method against published benchmarks on the BHSD dataset. All results report binary (foreground vs. background) Dice. The 2D methods operate on individual axial slices; the 3D methods process full volumetric sub-volumes. For completeness, we include the best-performing 3D supervised and semi-supervised results from the official BHSD benchmark [2]. Among 2D supervised methods, our Phase 1 multi-task model (0.747) already surpasses all single-task baselines, including the recent DE-UNeXt (0.704), by a margin of 4.3% absolute Dice. This gain stems from the joint training of segmentation and classification, which enriches encoder representations through lesion-guided attention. The semi-supervised Phase 2 further improves Dice to 0.779, outperforming the best 2D supervised baseline by 7.5% absolute and the best 3D semi-supervised method (CPS, 0.495) by a large margin.

The comparison across paradigms is indicative rather than strictly controlled. Although the 2D and 3D methods share the same evaluation target (binary segmentation), they differ in input formulation and training protocol, so the gap between the proposed 2D model and the 3D benchmarks should be interpreted

with caution; the within-paradigm 2D supervised comparison is the more reliable one.

4.6 Qualitative Analysis

We provide qualitative evidence for both tasks: segmentation outputs across our supervised and semi-supervised phases (Fig. 2), and classification attention maps comparing single-task and multi-task variants (Fig. 3).

Segmentation: Multi-Task vs. Semi-Supervised. Fig. 2 compares segmentation predictions from the Phase 1 multi-task model (Supervised) and the Phase 2 semi-supervised model (SSL) across all five ICH subtypes. The SSL model recovers missed hemorrhage fragments (IVH), suppresses false-positive detections (IPH, SDH), better delineates incomplete lesion boundaries (EDH), and produces more coherent masks for diffuse patterns (SAH). These observations are consistent with the quantitative gains in Table 4 and confirm that warm-start Mean Teacher regularization effectively leverages unlabeled data across diverse hemorrhage morphologies.

Classification: Effect of Lesion-Guided Attention. To assess whether the proposed lesion-guided mechanism improves classification reasoning, we compare Grad-CAM maps [19] from the single-task classifier and our multi-task model in Fig. 3. The single-task baseline often activates on broad or skull-adjacent regions, and in two cases incorrectly predicts EDH despite lesions belonging to other subtypes. In contrast, the multi-task model redirects attention toward the annotated hemorrhage regions and corrects both errors.

For correctly classified cases, lesion-guided attention also produces more compact and anatomically aligned activation maps, such as along the ventricular region for IVH and around localized extra-axial lesions for EDH/SDH. These results suggest that segmentation-derived spatial priors reduce shortcut reliance and help the classifier focus on clinically relevant hemorrhage morphology. Together with the quantitative F1 improvement in Tables 3–5, the Grad-CAM analysis supports the role of lesion-guided coupling in improving subtype recognition.

5 Conclusion

We presented a two-phase unified framework for joint intracranial hemorrhage segmentation and subtype classification from 2D CT slices. In Phase 1, a shared Wide-ResNet50-2 encoder jointly trains a UNet++ segmentation decoder and a novel lesion-guided classification head, exploiting the complementarity between pixel-level localization and image-level diagnosis. We demonstrated that the optimal input representation is a dual-window (brain+subdural) two-channel tensor, identified through systematic windowing ablation. In Phase 2, a warm-start

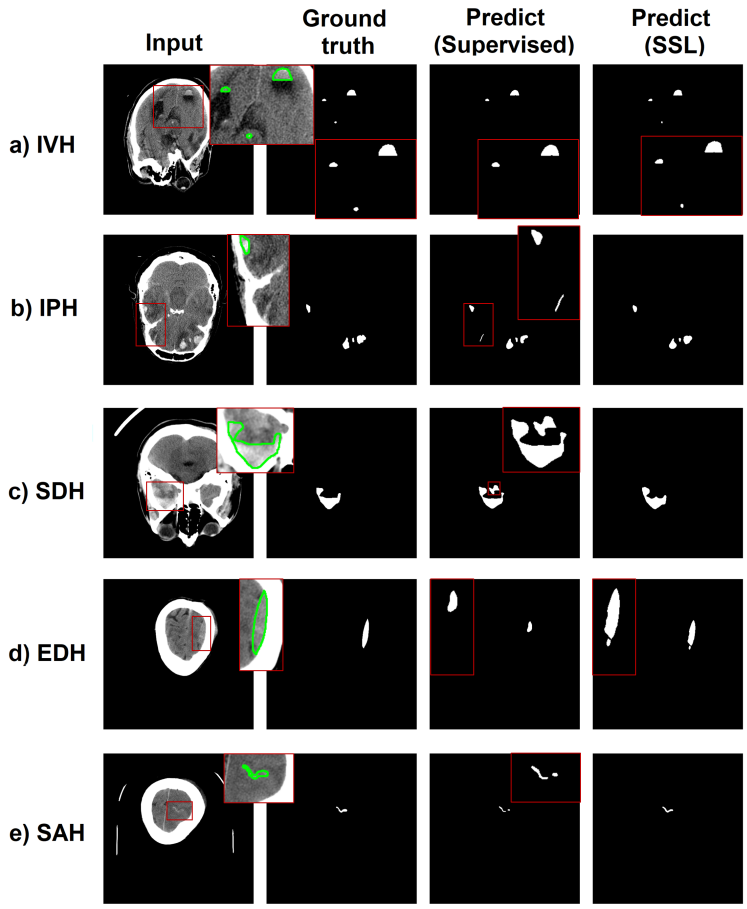

Fig. 2: Qualitative segmentation results on BHSD test set across all five ICH subtypes (a-e). Columns, from left to right: input CT slice with ground-truth hemorrhage contour overlaid in green; binary ground-truth mask (red box indicates the zoomed region shown in the inset); prediction from the Phase 1 (Supervised), and prediction from the Phase 2 (SSL). Red boxes in the prediction columns highlight problematic regions, typically false positives or missed lesion fragments, that are corrected or reduced by the SSL model.

Mean Teacher semi-supervised strategy leverages unlabeled CT slices through confidence-aware consistency regularization applied exclusively to the segmentation branch.

Experiments on BHSD confirm that multi-task learning substantially outperforms single-task baselines, and the semi-supervised stage further improves Dice to 0.779 and macro-F1 to 0.655. Grad-CAM visualizations further confirm that lesion-guided attention concentrates on hemorrhage regions more precisely than

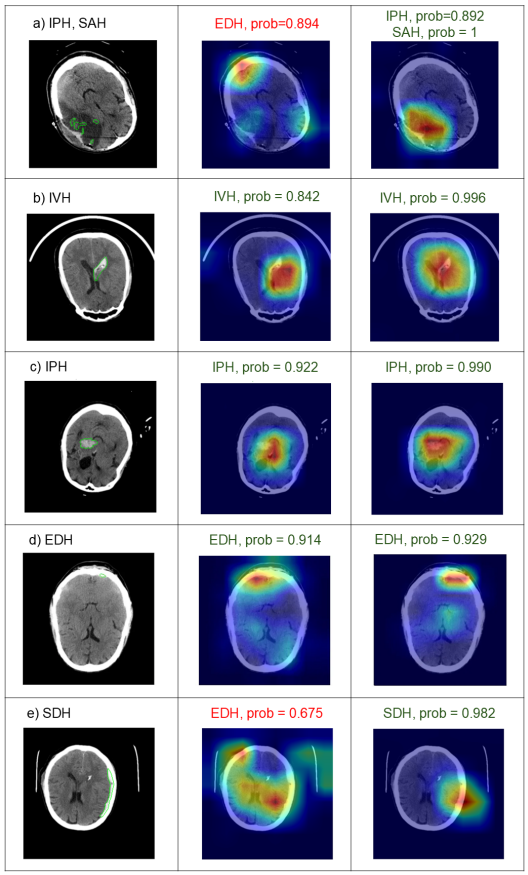

Fig. 3: Grad-CAM comparison between the classification-only baseline (**middle column**) and the proposed multi-task model (**right column**). The multi-task model produces more lesion-aligned attention and corrects representative baseline errors. The **left column** shows the input CT slice annotated with the ground-truth hemorrhage contour (green) and the ground-truth subtype label.

single-task baselines, and even corrects erroneous predictions when the single-task classifier relies on spurious skull-rim features. Future work will address the persistent under-performance on rare subtypes such as EDH, where extreme class imbalance limits classification gains despite improved spatial localization.

Acknowledgement. This research is funded by Hanoi University of Science and Technology (HUST) under project number T2025-PC-072.

References

1. Feigin, V.L., et al.: Global burden of stroke and risk factors in 188 countries. Lancet Neurol. **20**(10), 795–820 (2021)
2. Wu, B., et al.: BHSD: A 3D multi-class brain hemorrhage segmentation dataset. In: MLMI Workshop at MICCAI 2023, LNCS, vol. 14348, pp. 147–156. Springer (2023)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015, LNCS, vol. 9351, pp. 234–241. Springer (2015)
4. Zhou, Z., et al.: UNet++: A nested U-Net architecture for medical image segmentation. In: DLMIA 2018, LNCS, vol. 11045, pp. 3–11. Springer (2018)
5. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
6. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
7. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR, pp. 7482–7491 (2018)
8. Zhou, X., et al.: Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound. Med. Image Anal. **70**, 101925 (2021)
9. Mlynarski, P., et al.: Deep learning with mixed supervision for brain tumor segmentation. J. Med. Imaging **6**(3), 034002 (2019)
10. Lin, T.Y., et al.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)
11. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS, pp. 1195–1204 (2017)
12. Sohn, K., et al.: FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS, pp. 596–608 (2020)
13. Yu, L., et al.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: MICCAI 2019, LNCS, vol. 11765, pp. 605–613. Springer (2019)
14. Chan, T.: Computer aided detection of small acute intracranial hemorrhage on computer tomography of brain. Comput. Med. Imaging Graph. **31**(4–5), 285–298 (2007)
15. Chilamkurthy, S., et al.: Deep learning algorithms for detection of critical findings in head CT scans. Lancet **392**(10162), 2388–2396 (2018)
16. Valanarasu, J.M.J., Patel, V.M.: UNeXt: MLP-based rapid medical image segmentation network. In: MICCAI 2022, LNCS, vol. 13435, pp. 23–33. Springer (2022)
17. Chen, J., et al.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 (2021)
18. Wang, J., et al.: DE-UNeXt: Dual encoder UNeXt for intracranial hemorrhage segmentation on a novel HBU CH dataset. J. Radiat. Res. Appl. Sci. **18**, 101551 (2025)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)