# Weakly Supervised Intracranial Hemorrhage Segmentation with YOLO and an Uncertainty Rectified Segment Anything Model

Pascal Spiegler[1(✉)], Amirhossein Rasoulian[2], and Yiming Xiao[1]

[1] Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
`pascal.spiegler@mail.concordia.ca`
[2] NeuroRx Research, Montreal, Canada

**Abstract.** Intracranial hemorrhage (ICH) is a life-threatening condition that requires rapid and accurate diagnosis to improve treatment outcomes and patient survival rates. Recent advancements in supervised deep learning have greatly improved the analysis of medical images, but often rely on extensive datasets with high-quality annotations, which are costly, time-consuming, and require medical expertise to prepare. To mitigate the need for large amounts of expert-prepared segmentation data, we have developed a novel weakly supervised ICH segmentation method that utilizes the YOLO object detection model and an uncertainty-rectified Segment Anything Model (SAM). In addition, we have proposed a novel point prompt generator for this model to further improve segmentation results with YOLO-predicted bounding box prompts. Our approach achieved a high accuracy of 0.933 and an AUC of 0.796 in ICH detection, along with a mean Dice score of 0.629 for ICH segmentation, outperforming existing weakly supervised and popular supervised (UNet and Swin-UNETR) approaches. Overall, the proposed method provides a robust and accurate alternative to the more commonly used supervised techniques for ICH quantification without requiring refined segmentation ground truths during model training.

**Keywords:** Weak supervision · Image segmentation · Object detection · Medical imaging · Intracranial hemorrhage · YOLO · SAM

## 1 Introduction

Intracranial hemorrhage (ICH) accounts for 10–15% of all stroke cases and carries a significant risk of mortality [9]. Hemorrhage volume, which can rapidly expand within the first few hours, is a key predictor of treatment outcomes and potential complications [17]. Precise localization and quantification of the five ICH subtypes, including intraventricular (IVH), intraparenchymal (IPH), subarachnoid

(SAH), epidural (EDH), and subdural (SDH), are therefore essential for tailoring treatment strategies and minimizing adverse events [1]. While supervised deep learning (DL) models have demonstrated excellent potential in automating ICH assessment [7], their success heavily relies on large datasets with pixel-level annotations (ground-truth masks) and poor segmentation accuracy is observed with smaller training datasets [8]. However, large training datasets containing high-quality ground-truth masks are difficult to obtain due to high demands in time, labor, and domain expertise. Together with scarce public ICH segmentation datasets, this bottleneck poses great challenges in developing automatic ICH quantification algorithms to better facilitate the care and management of the condition.

To overcome the aforementioned issue, weakly supervised learning approaches [18,19] have emerged as a promising alternative. These methods leverage more economic ground truths, such as categorical labels, bounding boxes, or coarse masks to train segmentation models, bypassing the requirement of refined masks for fully supervised and semi-supervised approaches. While most existing literature is dedicated to ICH detection, ICH segmentation using weakly supervised methods remains under-explored. However, limited prior explorations exist leveraging explainable AI methods for weakly-supervised stroke segmentation, including class-activation maps (CAM) [25] and self-attention maps [19], providing encouraging results. Recent developments in foundation models, such as the Segment Anything Model (SAM) [10] have shown great potential to mitigate the segmentation ground truth bottleneck, but have not been explored for improving weakly supervised ICH segmentation. Therefore, we propose a novel weakly supervised ICH segmentation technique that incorporates automatic box and point prompt generation with SAM to allow for ICH detection and segmentation on CT scans. We have three main contributions. **First**, we leveraged a finetuned YOLOv8 model and a novel morphology-based method to automatically generate box and point prompts, respectively, for SAM. **Second**, to enhance segmentation accuracy with SAM, we employed an uncertainty rectification approach to account for uncertainty in prompt generation. **Lastly**, we explored the impacts of different prompt types for our proposed framework in ICH segmentation and compared it against state-of-the-art (SOTA) supervised and weakly supervised techniques.

## 2   Related Works

ICH segmentation methods still primarily rely on fully supervised approaches [2,3,11,12] and often with in-house datasets. More recently, semi-supervised techniques [24] have also been proposed for ICH quantification. However, refined segmentation ground truths are still crucial for their success, and more practical weakly supervised methods are gaining interest. In the limited prior works in this direction, most have relied on categorical labels as weak ground truths. For example, Wu et al. [25] proposed to use refined CAM results and representation learning to achieve ischemic stroke lesion segmentation, achieving a

0.3827 mean Dice score on multi-spectral MRIs. Later, from a binary classification CNN, Nemcek et al. [16] detected the location of ICH as bounding boxes in axial brain CT slices using the local extrema of derived attention maps, with a mean Dice of 0.58 for the lesion bounding boxes. Recently, Rasoulian et al. [19] utilized Head-Wise Gradient-Infused Self-Attention Maps from a Swin Transformer (Swin-HGI-SAM) trained on binary labels (ICH vs. no ICH) to obtain ICH segmentation, which obtained a mean Dice score of 0.438 on CT scans. The recent introduction of SAM [10], which allows interactive prompting in the forms of bounding boxes and/or points for zero-shot segmentation has attracted significant attention. However, its performance on CT-based ICH quantification and as an integrated solution allowing full automation in weakly supervised segmentation is yet to be explored. Furthermore, YOLO models [23] have been employed for ICH detection [4], but no reports have investigated their potential to facilitate the automation of SAM in ICH segmentation thus far.

## 3    Methods and Materials

### 3.1    Dataset and Preprocessing

For our study, we used the public Brain Hemorrhage Extended (BHX) dataset [20], which includes bounding box annotations for ICH along with their corresponding lesion subtypes, and the manually labeled PhysioNet CT dataset [8], which includes manual ICH segmentations. While 4607 CT slices and 5543 bounding boxes from the BHX dataset (containing the ICH subtypes and healthy scans) were employed to train and validate the YOLO model for lesion bounding box detection, the PhysioNet ICH segmentation dataset, which has 2814 CT slices with 318 mask-annotated ICH slices, was reserved as an independent test set to evaluate ICH segmentation with SAM. In addition, for our selected fully supervised baseline methods (more details in Sect. 3.3), subject-wise five-fold cross-validation was used on the manually segmented PhysioNet dataset to provide segmentation results for all cases and ensure that no slices from the same subject exist across different folds. As CT scans typically have a high dynamic range, for each CT slice, brain, subdural, and bone windows were created based on previous guidelines [5] and stacked together to form a composite RGB image, which was normalized to the range of [0,1] in each channel to facilitate training.

### 3.2    Uncertainty-Rectified YOLO-SAM Models

We propose YOLO-SAM, a novel weakly supervised framework for ICH segmentation, where the YOLOv8 model [23] provides several prompts for SAM to perform ICH segmentation. Here, we built three YOLO-SAM variants, including YOLO-SAM-BBox, YOLO-SAM-Point, and YOLO-SAM-PointBBox, which perform ICH segmentation using bounding box prompts, point prompts, and combinations of bounding boxes and point prompts, respectively. These models each employ an uncertainty rectification strategy that combines 10 SAM outputs
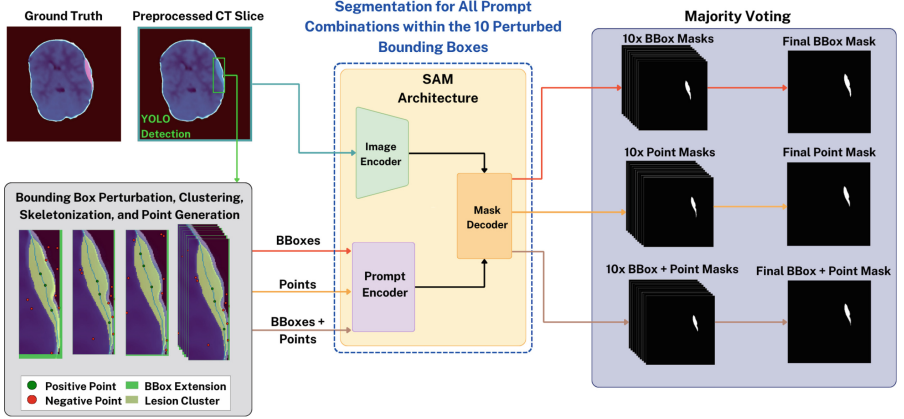
**Fig. 1.** Workflow of the proposed weakly supervised ICH segmentation method.

based on their 10 respective perturbed prompts. The detailed procedure of our methods is described below and shown in Fig. 1.

**YOLO Detection:** The preprocessed CT slices are passed to YOLOv8, which outputs the bounding boxes and associated lesion types for detected ICH. Then, the corner coordinates of the predicted bounding boxes are recorded to serve as the basis for automatic point prompt generation.

**Bounding Box Perturbations:** To enhance segmentation robustness and facilitate downstream uncertainty rectification in SAM's outputs, we introduce a method involving bounding box perturbation. Specifically, each YOLO-predicted bounding box is perturbed 10 times by randomly increasing its size by 1–4 pixels on each side. These perturbed boxes are recorded for the next step.

**Clustering and Point Prompt Generation:** Next, to strengthen the prompts' efficacy for SAM, leveraging the predicted ICH bounding box from YOLO, we introduce a novel point prompt generation method for the lesion and background based on a tailored tissue clustering solution and morphological analysis. To delineate lesions in proximity to the skull (e.g. SDH hemorrhage) for SAM, BET [22] skull-stripping is first applied to the entire CT image. Then, within the ICH bounding box for the skull-stripped RGB composite CT slice, K-means clustering is applied for tissue classification. Here, we use four clusters $(K = 4)$ regardless of hemorrhage sub-types. If any residual skull is present, in decreasing order of the Hounsfield unit (HU) value, we must account for 1) residual skull tissue 2) ICH 3) healthy brain tissue and 4) dark background; if not, we can expect the brightest cluster to be assigned to the lesion and the remaining 3 clusters to be assigned to the rest. Then, an algorithm is devised to automatically identify the lesion cluster out of the four (YOLO-Clustering). The resulting simple tissue clustering is obtained by first inspecting whether the cluster with the highest average HU value corresponds to the brightest signals

in the bone window channel, which represents the residual bone. If not, the cluster is selected as the lesion cluster. Otherwise, the algorithm picks the cluster with the second-highest average HU value. Finally, on the K-means-based lesion clusters, skeletonization is performed to extract the skeleton of the shapes. From these skeletons, positive ICH point prompts are sampled. Then, from each of the three other clusters, negative points are sampled for SAM segmentation.

**SAM Segmentation with Uncertainty Rectification:** For each of the 10 perturbed bounding boxes, each combination of generated prompts (bounding box, points, and point-box) are passed to SAM's prompt encoder along with the input image to produce a segmentation sample (Fig. 1). For each YOLO-SAM variant, their final segmentation is obtained via majority voting based on 10 segmentation samples from the associated prompt type. This voting mechanism ensures the robustness of ICH segmentation against network-related prompt instability and SAM's potential sensitivity to these variations, further improving segmentation quality.

### 3.3   Baseline Models and Ablation Study

To validate our proposed method, we compared its performance against the SOTA weakly supervised and fully supervised segmentation techniques for ICH segmentation. With an open-source repository and good performance, we chose the recent Swin-HGI-SAM [19] as our weakly supervised baseline. In terms of baseline methods with full supervision, we selected the popular UNet [21] and Swin-UNETR models [6], which have demonstrated strong performance in a wide range of medical image segmentation tasks. For the UNet model, we implemented the architecture from the manually segmented PhysioNet CT data paper [8], with four hierarchical layers in the encoding and decoding paths. For the Swin-UNETR model [6], we also adopted four hierarchical levels to be consistent with the UNet model.

While the SAM model [10] allows both bounding boxes and/or points as interactive prompts to generate segmentation results, the robustness and accuracy of individual prompt types and their combined usage still require further investigation. Therefore, besides comparison with the baseline models, we also performed an ablation study on the impact of prompt types for the target task (YOLO-SAM-Point, YOLO-SAM-BBox, YOLO-SAM-PointBBox).

### 3.4   Model Training and Evaluation Metrics

The YOLOv8-m model pretrained on the MS COCO dataset [13] in our YOLO-SAM variants was finetuned on the BHX dataset [20], with 3685 CT-slice images and 4479 bounding box labels for the training set, as well as 922 CT-slice images with 1064 labels for the validation set. We used the default YOLOv8 configuration (batch size=16, patience=100) during training. For the first 10,000 iterations, the AdamW optimizer was used with a learning rate of 0.00111 (calculated by a fitting equation using the number of bbox classes, which was 5 for

each ICH subtype) and a momentum of 0.9. For the remaining iterations (after epoch 44), the SGD optimizer with an initial learning rate of 0.01 and momentum of 0.9 was used. We trained the weakly supervised Swin-HGI-SAM model [19] with the RSNA 2019 Brain CT hemorrhage dataset [5] (90%:10% data split for training vs. validation) following the details from the original publication. As for the supervised baselines (UNet and Swin-UNETR), subject-wise five-fold cross-validation was employed exclusively on the manually segmented PhysioNet dataset, using the AdamW optimizer with an initial learning rate of 0.001 as well as a loss function based on Dice coefficient and cross-entropy. We conducted all model training on a desktop computer with an Intel Core i9 CPU and an NVIDIA GeForce RTX 3090 GPU. *After model training, all evaluations were based on the manually segmented PhysioNet dataset in a slice-wise manner using the default YOLO confidence threshold of 0.25.* As ICH detection is a crucial component of our method, besides segmentation, we also evaluated the binary ICH detection performance (ICH vs. no ICH) for all DL models with accuracy, precision, recall, AUC, F1-score, and specificity. Note that a YOLO prediction was considered a true positive if it correctly identified a slice containing ICH, irrespective of the predicted subtype. For UNet and Swin-UNETR, a true-positive detection was defined as a slice with ICH segmentation that contains more than 10 pixels. In terms of segmentation, we computed the Dice coefficient and Intersection over Union (IoU) for all proposed and baseline models. Paired two-sample t-tests were then used to compare the Dice and IoU scores between the proposed method and the baselines, with $p < 0.05$ indicating a statistically significant difference.

## 4 Results

### 4.1 Detection Performance

**Table 1.** Detection Performance of Different Methods

| Metric | Swin-HGI-SAM | U-Net | Swin-UNETR | YOLOv8-m |
|---|---|---|---|---|
| Accuracy | 0.950 | 0.647 | 0.655 | 0.933 |
| Precision | 0.765 | 0.239 | 0.253 | 0.665 |
| Recall | 0.791 | 0.901 | 0.907 | 0.626 |
| AUC | 0.880 | 0.757 | 0.764 | 0.796 |
| F1-score | 0.767 | 0.373 | 0.374 | 0.645 |
| Specificity | 0.969 | 0.612 | 0.622 | 0.966 |

The ICH detection performance for all models is listed in Table 1. Similar to Swin-HGI-SAM, the YOLOv8-m model demonstrated superior detection performance for most metrics compared to the U-Net and Swin-UNETR models, particularly in precision (0.665 vs. 0.239 and 0.253), AUC (0.796 vs. 0.757 and 0.764), F1-score (0.645 vs. 0.373 and 0.374), and specificity (0.966 vs. 0.612 and

0.622). This highlights the potential of using bounding box localization models such as YOLO to achieve superior performance compared to mask-trained approaches on limited data (UNet, Swin-UNETR) and competitive performance with models trained on substantially more binary labels (Swin-HGI-SAM). However, a weakness of the YOLOv8-m model is its lower slice-wise recall compared to Swin-HGI-SAM (0.626 vs. 0.791), indicating that Swin-HGI-SAM will more reliably detect true positives. For all other detection metrics, YOLOv8-m demonstrated comparable but marginally weaker detection performance, likely due to the smaller number of training samples (4607 bounding box annotated CT slices for YOLO versus 677523 binary-labelled CT slices for Swin-HGI-SAM).

## 4.2   Segmentation Performance

The ICH segmentation results are shown in Table 2, with qualitative outcomes demonstrated in Fig. 2. Table 2 shows that point prompts, hybrid point and bounding box prompts, as well as simple tissue clustering within the YOLO bounding box (YOLO-Clustering) yielded significantly higher segmentation performance than Swin-HGI-SAM, UNet, Swin-UNETR and using bounding box prompts alone ($p < 0.005$). It is also shown that hybrid prompts have improved performance over point prompts and YOLO-Clustering on average, though not statistically significant ($p > 0.05$). While YOLO-Clustering had good segmentation quality, it also had higher standard error than SAM with hybrid and point prompts, highlighting the point prompt's better precision and reliability. Finally, while YOLO-SAM-BBox does not show significantly higher Dice and IoU scores than UNet ($p = 0.0853$) or Swin-UNETR ($p = 0.768$), it significantly outperforms Swin-HGI-SAM ($p < 0.005$).
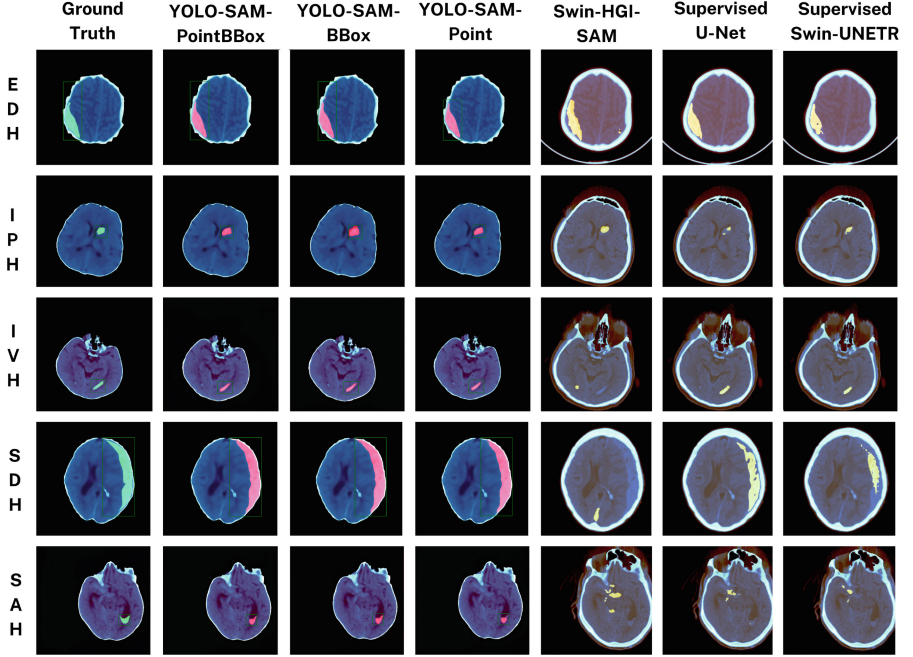
**Table 2.** Segmentation Performance of Different Models (mean ± standard error)

| Model | Dice | IoU |
|---|---|---|
| Swin-HGI-SAM | $0.403 \pm 0.014$ | $0.283 \pm 0.011$ |
| Fully supervised U-Net | $0.388 \pm 0.019$ | $0.297 \pm 0.016$ |
| Fully supervised Swin-UNETR | $0.428 \pm 0.018$ | $0.330 \pm 0.011$ |
| YOLO-Clustering | $0.625 \pm 0.020$ | $0.506 \pm 0.019$ |
| YOLO-SAM-BBox | $0.562 \pm 0.020$ | $0.445 \pm 0.018$ |
| **YOLO-SAM-Point** | $\mathbf{0.627 \pm 0.018}$ | $\mathbf{0.506 \pm 0.017}$ |
| **YOLO-SAM-PointBBox** | $\mathbf{0.629 \pm 0.018}$ | $\mathbf{0.508 \pm 0.017}$ |

## 5   Discussion

Our YOLO-SAM framework that integrates YOLOv8-m, a novel point-prompt generator, and SAM with uncertainty rectification has demonstrated great performance in weakly supervised ICH segmentation, particularly with the hybrid

**Fig. 2.** Qualitative segmentation results on different ICH subtypes

prompts. The superior performance over existing weakly supervised and fully supervised methods can be explained by the incorporation of the power of the foundation models and spatial information represented by the bounding box ground truths. It is important to acknowledge that the poor performance of fully supervised DL models, such as UNet and Swin-UNETR can also be partially due to the low number of ground-truth mask labels. Despite this success, the slice-wise recall metric for our YOLO model lagged behind the Swin-HGI-SAM, suggesting a potential compromise in the model's ability to detect all ICH slices. However, after investigating this further on a patient-wise basis, the recall metric was calculated at 0.9714, with 34 out of 35 patients with hemorrhage having had at least one slice detected. In a clinical setting, the proportion of true positive ICH cases would therefore be much higher than the reported slice-wise recall metric. Our ablation study showed that hybrid prompts offered better performance than points or bounding boxes. This observation echoes previous reports [15] and could be explained by the lack of robustness when capturing thin, elongated, and curved structures (e.g., IPH subtype) with bounding boxes by SAM. Finally, while MedSAM [15] has gained great popularity in the community, its adoption in our YOLO-MedSAM-BBox model resulted in inferior segmentation outcomes (Dice = 0.412 ±0.018, IoU = 0.298 ±015). This is consistent with other reports of SAM outperforming MedSAM on certain medical image segmentation tasks [14] and may be due to a lack of public datasets for

training MedSAM on ICH tasks, as Dice loss was used in training the model [15].

## 6    Conclusion

In conclusion, we have proposed a novel weakly supervised ICH segmentation technique that uses YOLO and an uncertainty-rectified SAM. In addition to bounding boxes provided via YOLO, our morphology-based point prompt generation was proven to offer enhanced segmentation performance. Thorough assessments have revealed its superior performance over SOTA weakly supervised and fully supervised baselines while maintaining strong ICH detection capabilities.

**Disclosure of Interests.** The authors declare no competing interests.

## References

1. Aguilar, M.I., Brott, T.G.: Update in intracerebral hemorrhage. Neurohospitalist **1**(3), 148–159 (2011)
2. Chang, P.D., et al.: Hybrid 3d/2d convolutional neural network for hemorrhage evaluation on head ct. Am. J. Neuroradiol. **39**(9), 1609–1616 (2018)
3. Cho, J., et al.: Improving sensitivity on identification and delineation of intracranial hemorrhage lesion using cascaded deep learning models. J. Digit. Imaging **32**, 450–461 (2019)
4. Ertuğrul, Ö.F., Akıl, M.F.: Detecting hemorrhage types and bounding box of hemorrhage by deep learning. Biomed. Signal Process. Control **71**, 103085 (2022)
5. Flanders, A.E., et al.: Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. Radiology: Artifi. Intell. **2**(3), e190211 (2020)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop, pp. 272–284. Springer (2021). https://doi.org/10.1007/978-3-031-08999-2_22
7. Heit, J., et al.: Automated cerebral hemorrhage detection using rapid. Am. J. Neuroradiol. **42**(2), 273–278 (2021)
8. Hssayeni, M.D., Croock, M.S., Salman, A.D., Al-khafaji, H.F., Yahya, Z.A., Ghoraani, B.: Intracranial hemorrhage segmentation using a deep convolutional model. Data **5**(1) (2020)
9. Isabel C. Hostettler, D.J.S., Werring, D.J.: Intracerebral hemorrhage: an update on diagnosis and treatment. Expert Rev. Neurotherapeutics **19**(7), 679–694 (2019), pMID: 31188036
10. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

11. Kuo, W., Häne, C., Yuh, E., Mukherjee, P., Malik, J.: Cost-sensitive active learning for intracranial hemorrhage detection. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 715–723. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_82

12. Lee, H., et al.: An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat. Biomed. Eng. **3**(3), 173–182 (2019)

13. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

14. Liu, H., Yang, H., van Diest, P.J., Pluim, J.P., Veta, M.: Wsi-sam: multiresolution segment anything model (sam) for histopathology whole-slide images. arXiv preprint arXiv:2403.09257 (2024)

15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)

16. Nemcek, J., Vicar, T., Jakubicek, R.: Weakly supervised deep learning-based intracranial hemorrhage localization. arXiv preprint arXiv:2105.00781 (2021)

17. Qureshi, A., Palesch, Y.: Antihypertensive treatment of acute cerebral hemorrhage (atach) ii: design, methods, and rationale. Neurocrit. Care **15**, 559–576 (2011)

18. Rasoulian, A., Salari, S., Xiao, Y.: Weakly supervised intracranial hemorrhage segmentation using hierarchical combination of attention maps from a swin transformer. In: Machine Learning in Clinical Neuroimaging, pp. 63–72. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-17899-3_7

19. Rasoulian, A., Salari, S., Xiao, Y.: Weakly Supervised Intracranial Hemorrhage Segmentation using Head-Wise Gradient-Infused Self-Attention Maps from a Swin Transformer in Categorical Learning. arXiv (Cornell University) (Jan 2023)

20. Reis, E., et al.: Brain hemorrhage extended (bhx): bounding box extrapolation from thick to thin slice ct images (July 2020)

21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

22. Smith, S.M.: Fast robust automated brain extraction. Hum. Brain Mapp. **17**(3), 143–155 (2002)

23. Terven, J., Córdova-Esparza, D.M., Romero-González, J.A.: A comprehensive review of yolo architectures in computer vision: from yolov1 to yolov8 and yolo-nas. Mach. Learn. Knowl. Extract. **5**(4), 1680–1716 (2023)

24. Wang, J.L., Farooq, H., Zhuang, H., Ibrahim, A.K.: Segmentation of intracranial hemorrhage using semi-supervised multi-task attention-based u-net. Appl. Sci. **10**(9), 3297 (2020)

25. Wu, K., Du, B., Luo, M., Wen, H., Shen, Y., Feng, J.: Weakly supervised brain lesion segmentation via attentional representation learning. In: Shen, D., et al (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 211–219. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_24