Article

Patterns

Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network

Highlights

- GLIA-Net is a deep learning method for the clinical diagnosis of IAs
- It can be applied directly to CTA images without any laborious preprocessing
- A clinical study demonstrates its effectiveness in assisting diagnosis
- An IA dataset of 1,338 CTA cases from six institutions is publicly released

Authors

Zi-Hao Bo, Hui Qiao, Chong Tian, ..., Tijiang Zhang, Rongpin Wang, Qionghai Dai

Correspondence

feng-xu@tsinghua.edu.cn (F.X.), tijzhang@163.com (T.Z.), wangrongpin@126.com (R.W.), qhdai@tsinghua.edu.cn (Q.D.)

In Brief

Intracranial aneurysm (IA) diagnosis on CTA images is laborious and time consuming in clinical routine. By combining global risk prediction with local IA recognition, the proposed GLIA-Net can robustly and efficiently assist radiologists in clinical practice without any pre- or postprocessing. The state-of-the-art performance has been validated via a multi-cohort dataset, which has been publicly released to democratize deep learning algorithms for biomedical research.

Bo et al., 2021, Patterns 2, 100197
February 12, 2021
https://doi.org/10.1016/j.patter.2020.100197

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network

Zi-Hao Bo,[1,9] Hui Qiao,[2,4,9] Chong Tian,[3,9] Yuchen Guo,[2] Wuchao Li,[3] Tiantian Liang,[3] Dongxue Li,[3] Dan Liao,[3] Xianchun Zeng,[3] Leilei Mei,[5] Tianliang Shi,[6] Bo Wu,[6] Chao Huang,[6] Lu Liu,[7] Can Jin,[7] Qiping Guo,[8] Jun-Hai Yong,[1] Feng Xu,[1,4,10,*] Tijiang Zhang,[5,*] Rongpin Wang,[3,*] and Qionghai Dai[2,4,*]

[1]BNRist and School of Software, Tsinghua University, Beijing, Beijing 100084, China
[2]BNRist and Department of Automation, Tsinghua University, Beijing, Beijing 100084, China
[3]Department of Radiology and Guizhou Provincial Key Laboratory of Intelligent Medical Image Analysis and Precision Diagnosis, Guizhou Provincial People's Hospital, Guiyang, Guizhou 550002, China
[4]Institute of Brain and Cognitive Sciences, Tsinghua University, Beijing, Beijing 100084, China
[5]Department of Radiology, Affiliated Hospital of Zunyi Medical University, Zunyi, Guizhou 563000, China
[6]Department of Radiology, Tongren Municipal People's Hospital, Tongren, Guizhou 554300, China
[7]Department of Radiology, The Second People's Hospital of Guiyang, Guiyang, Guizhou 550002, China
[8]Department of Radiology, Xingyi Municipal People's Hospital, Xingyi, Guizhou 562400, China
[9]These authors contributed equally
[10]Lead contact
*Correspondence: feng-xu@tsinghua.edu.cn (F.X.), tijzhang@163.com (T.Z.), wangrongpin@126.com (R.W.), qhdai@tsinghua.edu.cn (Q.D.)
https://doi.org/10.1016/j.patter.2020.100197

THE BIGGER PICTURE Intracranial aneurysms (IAs) are enormous threats to human health with a prevalence of approximately 4%. The rupture of IAs usually causes death or severe damage to the patients. To enhance the clinical diagnosis of IAs, we present a deep learning model (GLIA-Net) for IA detection and segmentation without laborious human intervention, which achieves superior diagnostic performance validated by quantitative evaluations as well as a sophisticated clinical study. We anticipate that the publicly released data and the artificial intelligence technique would help to transform the clinical diagnostics and precision treatments of cerebrovascular diseases. They may also revolutionize the landscape of healthcare and biomedical research in the future.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Intracranial aneurysm (IA) is an enormous threat to human health, which often results in nontraumatic subarachnoid hemorrhage or dismal prognosis. Diagnosing IAs on commonly used computed tomographic angiography (CTA) examinations remains laborious and time consuming, leading to error-prone results in clinical practice, especially for small targets. In this study, we propose a fully automatic deep-learning model for IA segmentation that can be applied to CTA images. Our model, called Global Localization-based IA Network (GLIA-Net), can incorporate the global localization prior and generates the fine-grain three-dimensional segmentation. GLIA-Net is trained and evaluated on a big internal dataset (1,338 scans from six institutions) and two external datasets. Evaluations show that our model exhibits good tolerance to different settings and achieves superior performance to other models. A clinical experiment further demonstrates the clinical utility of our technique, which helps radiologists in the diagnosis of IAs.

INTRODUCTION

The diagnosis and treatment of intracranial aneurysms (IAs) are important and difficult in clinical assessment. As reported, the prevalence of IAs in the general population can be up to 6%,[1] and the rupture of IAs is always associated with severe morbidity and mortality.[2] According to statistics, 80% of nontraumatic subarachnoid hemorrhages are caused by the rupture of IAs,[3,4] and most of the nontraumatic subarachnoid hemorrhages will result in death or dismal prognosis.[5,6] However, the small size of IAs and low-intensity contrast to normal vessels in medical image scans makes even subspecialty-trained radiologists need to

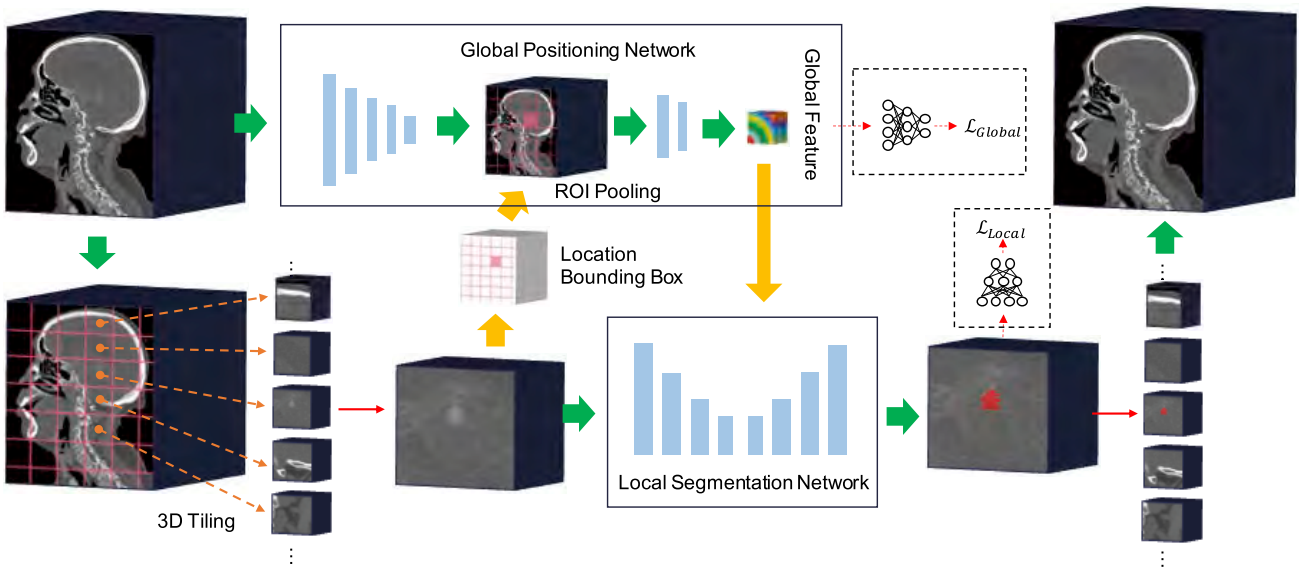


Figure 1. The workflow of our GLIA-Net model in IA segmentation

The original CTA image is directly used as the global image. A 3D tiling method is adopted and generates many local images. The model consists of two parts: (1) global positioning network, which analyzes the global image and gives a risk distribution map roughly to the (2) local segmentation network, which uses the local images to generate voxel-wise segmentation results. The final segmentation map is constructed from all the local image patches.

check up to hundreds of image slices carefully for one patient. Moreover, individual radiologists may not make consistent assessments, also emphasizing the diagnostic difficulties. Aiming to improve the diagnosis, delicately developed algorithms for the detection of IAs from computed tomography angiography (CTA) and magnetic resonance angiography (MRA) have been proposed over the past several decades.[7–10] Although digital subtraction angiography (DSA) is still the gold standard for diagnosing IAs, CTA has been proved to be able to diagnose IAs in most situations[11–14] and is a non-invasive, time-saving, and cost-effective technique with usually wider availability.[15] Besides, compared with MRA, CTA is a faster, widely available diagnostic technique that costs less and has higher image resolution.

The essential role of clinical diagnosis is to detect IAs and generate three-dimensional (3D) segmentation. Despite the rapid development of biomedicine, radiologists still need to make great efforts in the manual detection of IA lesions from medical images, and with highly limited computer assistance. Conventional automatic or semi-automatic segmentation methods for IAs have been proposed during the past few years.[16,17] However, they are quite sensitive to different device settings and require a lot of laborious pre and post operations.

Offering an alternative solution, recent advances in deep learning have shown great potential for medical image interpretation, which promises to assist radiologists and clinicians to speed up and improve the clinical diagnosis. Inspired by the development of deep learning on segmentation for natural image interpretation, several deep neural networks, such as U-Net,[18] V-net,[19] 3D U-Net,[20] and P-Net,[21] have been validated successfully on biomedical datasets recently. Such successes stimulate an upsurge of research interest in deep-learning-based biomedical applications, including the diagnosis of IAs.[7,8] Nevertheless, current deep learning models for IA image interpretation seldom consider the inherent characteristics of biomedical data and still need human interventions like brain extraction or bone removal, which highly depend on the human and device resources in different institutions. Besides, the training datasets of these models are pretty small, comprising no more than a few hundreds of aneurysms. Given that the training data are usually collected from a single institution, it is hard to guarantee the generalization capability of the learned models when facing a large, multi-institutional cohort.

To take a step further in the diagnosis of IAs, we first investigated all the difficulties in achieving IA segmentation from medical images. The difficulties are listed as follows. Medical images for IA diagnosis are typically quite large (over 500 pixels in all the three dimensions) and it is difficult to identify the 3D structural lesion regions from just a single 2D image. Besides, patients may undergo thorough examinations that cover not only the head region but also the neck or lung/heart region, which increases the size of the original CTA scans. However, processing 3D images directly has been impossible for such a large image size so far because of the limited computation resources. Besides, the size of IAs is generally small compared with the whole image, making it hardly possible to down-sample the image to meet the computation resource requirement as the aneurysms may be lost.

There are some methods that were designed to deal with the huge amount of data in medical images. Some methods directly use 2D or 2.5D segmentation methods.[22–24] Kong et al.[25] used a recurrent neural network (RNN) to encode information among several 2D slices, but these methods only work well for large object segmentation, such as cardiac and kidney partition. As a compromise between 3D feature requirement and limited computation resource, some methods use the 3D patching strategy, which clips a small cube as an input sample.[20,26] Such 3D methods can extract local shape information very well but raise another problem: the global structure information is lost in clipped patches. This can be ignored in some applications like tumor

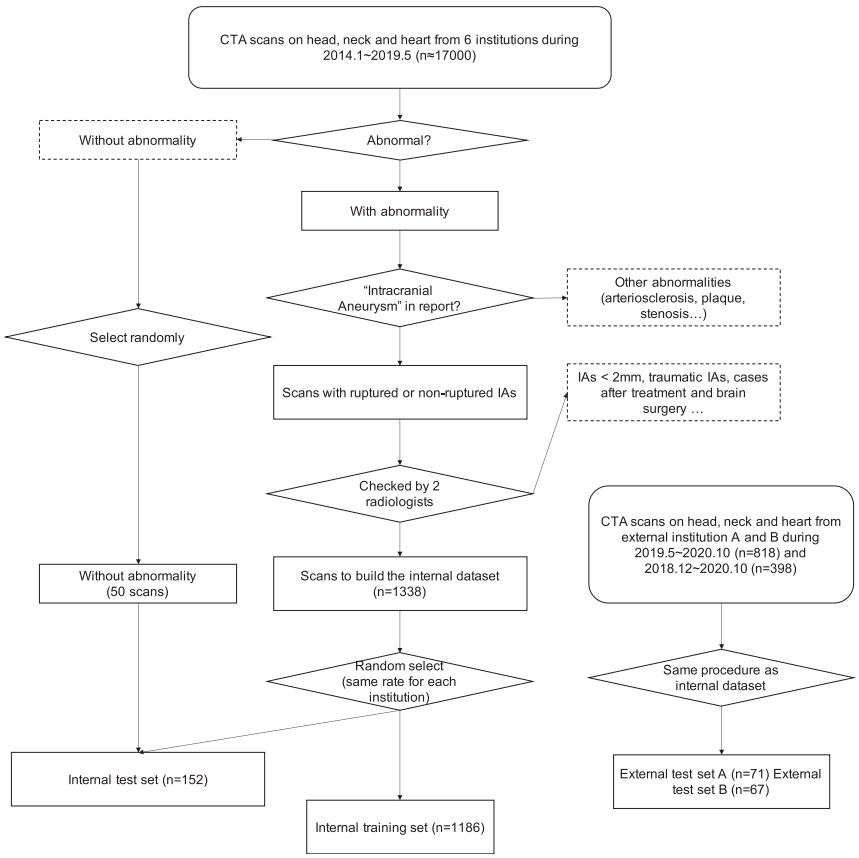


Figure 2. Flowchart of data selection and division

The n in the flowchart means the number of cases.

detection in pathology slices but is very important for the lesions that have a global position tendency in the body, like IAs. Although there are some works considering that multi-level feature fusion can achieve a plausible effect on targets with different sizes,[27,28] they still adopt the whole image as input, which means they cannot accommodate such huge CTA images in one glance. Some medical imaging models do consider the location information,[29–31] but they just directly use the coordinate values without delicately extracting and representing the location information, thus they are closely attached to the data they use.

Based on these considerations, in this paper, we propose a strategy to adopt global structure information in the normal 3D patching segmentation network. Our model is called the Global Localization-based IA Network (GLIA-Net). The overview of our model is shown in Figure 1. This method can be used in any 3D sliding-cube convolution tasks, especially for those requiring strong global position information. Besides, we propose a strategy to design spatial variant losses for annotated pixels. This is to consider the annotation inconsistency on the lesion edges.[32] We use CTA images because they have a short examination period in emergency treatment and are more economical to obtain in many countries. The segmentation model in this work does not need any pre- or postprocessing procedures with human intervention, like brain extraction or bone removal, and can also work on CTA images containing neck or lung regions without any advanced accessory equipment or medical software. The internal dataset in this work contains 1,338 CTA images with 1,489 IAs from six different institutions, which guarantees the generalization ability of the model. Compared with the existing deep learning models, GLIA-Net achieves a target-wise recall rate of 82.1 (78.2–86.0, 95% confidence interval [CI]) on the internal test set with 4.38 (2.91–5.85, 95% CI) false-positive IAs per case, and a voxel-wise segmentation average precision (AP) of 61.9 (59.4–64.4, 95% CI), showing superior diagnostic performance. To further validate the clinical feasibility of our model, a clinical experiment has also been conducted.

RESULTS

Data

The large variety of IAs' positions, shapes, and sizes makes them hard to identify. Thus, a large dataset with different but complementary aspects of patients is required for training and evaluation. We collected an internal dataset from six institutions (Guizhou Provincial People's Hospital, Affiliated Hospital of Zunyi Medical University, Tongren Municipal People's Hospital, Xingyi Municipal People's Hospital, The Second People's Hospital of Guiyang, The First People's Hospital of Zunyi) and two external datasets from two institutions (People's Hospital of Anshun City, Zhijin People's Hospital). The internal dataset contains 1,338 CTA images with 1,489 IAs, which contributes to a huge number of 699,266 512 × 512 2D image slides in total (approximately 174GB of disk space). The pipeline to build the dataset and the inclusion and exclusion criteria are shown in Figure 2.

The images all contain the head region of the patients, some of which may also contain the neck or heart region. The data include non-ruptured ones and ruptured ones with subarachnoid hemorrhage or parenchymal hemorrhage. The CTA scans in the internal dataset were captured by 11 devices that belong to six equipment models (SIEMENS SOMATOM Definition AS+, SIEMENS SOMATOM Definition Flash, SIEMENS SOMATOM Force, NMS NeuViz 128, GE MEDICAL SYSTEMS Discovery CT, SIEMENS Sensation 64) from three manufactures (Siemens, Neusoft, and GE Healthcare). There are four different scan layer thicknesses (0.6 mm, 0.625 mm, 0.75 mm, and 1.0 mm). All the patients were in the head-first-supine (HFS) position during the examination with peak voltage between 70 and 140 kV, and tube current between 45 and 1,275 mA.

The CTA images were annotated by five clinicians and reviewed by two experienced CTA diagnosis radiologists with the associated clinical and personal information of the patients. The identified IAs were manually segmented on each slice by using the open-source annotation software ITK-SNAP.[33] The

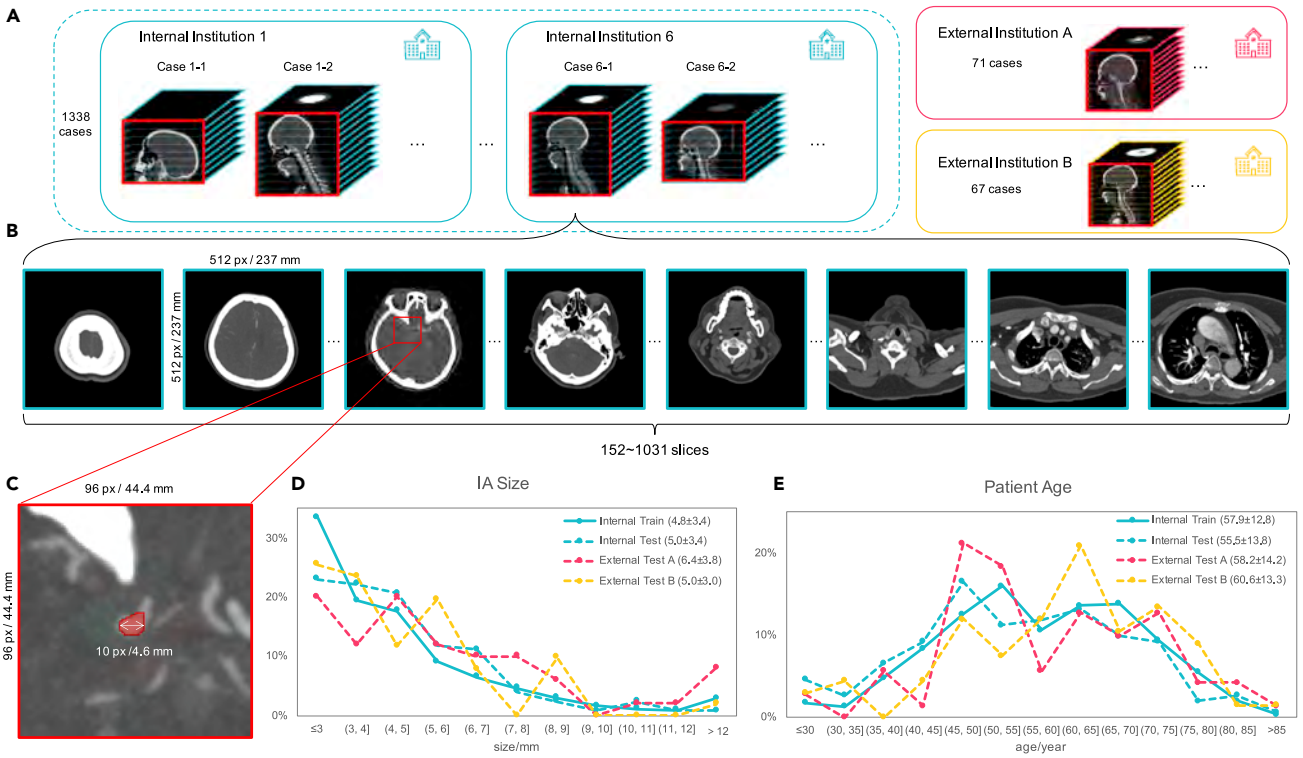


Figure 3. Dataset illustration in this study

The IAs are typically tiny in the whole CTA images (about 1/1,000,000 in voxels), which makes the diagnosis a tough task.

(A) Our internal dataset contains 1,338 CTA cases (1,489 IAs) that were collected from six different institutions and the two external test set contains 71 and 67 cases (50 and 51 IAs) respectively. The CTA scans used in training were captured by 11 devices that belong to seven equipment models coming from three manufactures with four different scan layer thicknesses.

(B) Each case contains 152–1,310 image slices with a size of 512 × 512 resulting in a total of 699,266 slices, which may include neck and heart regions associated with the head region as they were collected from real clinical examinations.

(C–E) (C) IAs are typically tiny compared with the whole CTA scans. The presence of IAs was annotated by five clinicians and reviewed by two experienced CTA diagnosis radiologists with reference to the patients' clinical reports and personal information. We also show the histograms of (D) patient age distribution and (E) IA size distribution for the dataset, and the legend also shows the mean and SD.

annotation was used as the ground truth standard both in training and evaluation, although it should be noted that there might be some bias or noise considering the vague boundary of IAs in CTA images and high inter-observer variability, which means they may not share the same labeling standard subconsciously.

The whole internal dataset was split into 1,186 cases for training and 152 cases for testing. The internal test set contains 50 negative cases (no IAs occur). We do not include negative cases in the training set, because the training already suffers from severe data imbalance as the IAs are small in the brain. We verified that all the positive CTA images distribute roughly equally for different institutions, ages, and genders in the internal training and test set. The training set was used to train the model and the test set was only used to evaluate the performance of our model, which means that the model could not see the images in the test set before the training is done.

To evaluate the generalization of our model, we also collected two external datasets from another two institutions that were not included in the internal dataset. The external test set A contains 71 (including 24 negatives) cases and the external test set B contains 67 (including 22 negatives) cases. The building process of these external datasets was the same as that in the internal dataset. The scans in the external dataset A were captured using one device, a Philips Ingenuity CT model, while those in the external dataset B were also captured using one device, a Philips Brilliance 64 model. These models were not included in the internal dataset. Slice thicknesses for the two external test sets are both 0.9 mm, which does not appear in the internal dataset either. The detailed dataset statistics can be viewed in Figure 3 and Table 1.

Performance evaluation

Our model uses a patching strategy to segment IAs across the whole original CTA scan (the global image) without bone removal or any other preprocessing. One CTA image is split into lots of small patches (the local image) and the network is applied on each patch at a time. At last, the results of all patches are combined to build the entire segmentation map. The segmentation network gives each voxel in the input 3D CTA patch a label that indicates whether it belongs to an IA or normal tissue. Our GLIA-Net consists of two components: (1) the global positioning network, which extracts the information of IA distribution in the global image; and (2) the local segmentation network, which is used to segment voxel-wise label in the current local image.

Table 1. Dataset statistics in detail

Dataset	No. of cases	No. of IAs	No. of cases that contain		Gender		No. of cases containing			
			Ruptured IAs	Non-ruptured IAs	Male	Female	0 IA	1 IA	2 IAs	≥3 IAs
Internal training	1,186	1,363	474	712	508	678	0	1,043	119	24
Internal test	152	126	42	60	63	89	50	85	13	4
External test A	71	50	29	18	32	39	24	44	3	0
External test B	67	51	25	20	33	34	22	40	4	1

The input to the local segmentation network keeps the original resolution, while that to the global positioning network is down-sampled to fit in the limited Graphics Processing Unit (GPU) memory. The major difference between our IA segmentation model and other segmentation models is that we use a global positioning network to equip our patch-based local segmentation network with global information that a normal segmentation model cannot access. The global information is not just conveyed by some simple features like the position coordinate, but with a deep neural feature from the global image. In the following, we will discuss the ability and performance of our model.

Considering most IAs are small compared with the whole brain area and our segmentation model needs to give prediction labels for each voxel in the image, which is harder than just a binary patient classification objective, we used multiple related metrics to evaluate the performance of our model. To begin with, the AUC (area under curve) value computed on the ROC (receiver operating characteristic) curve[34] and the AP value computed on the precision-recall curve[35] are considered to demonstrate the overall performance of a segmentation task, which considers all the probability thresholds on the output. Then, we set a fixed probability threshold of 0.5 in the following qualitative evaluations. The precision, recall, DSC (dice similarity coefficient),[36] and 95% HD (the 95th percentile Hausdorff distance)[37] values were calculated in this threshold. To show the stability of the tested models, we also show the 95% confidence interval (95% CI) of all evaluation metrics, which was summarized from five different runs for each experiment setting. The performance can be explained in two aspects: voxel-wise segmentation performance and target-wise detection performance (see Figure 4).

Our segmentation model was trained on the training set of our internal dataset (see also in the Experimental procedures and the Supplemental experimental procedures) and was tested on the internal test set and two external test sets.

We also compared our GLIA-Net with two deep learning models. First, we compared our GLIA-Net with the enhanced version of U-Net,[18] which is also the baseline network of our local segmentation network. This deep learning network has been widely used in many segmentation tasks in the past few years and has shown great power compared with traditional models. In our experiment, we modified the original U-Net to a 3D version and replaced its original convolutions with residual blocks. Then, we compared our model with a state-of-the-art IA segmentation network called HeadXNet.[8] This model can be regarded as the ultimate version of U-Net supported by the advanced feature extracting network SE-ResNeXt[38,39] and atrous spatial pyramid pooling.[40] We chose this network as a comparison because it showed great potential in IA segmentation and its ability was tested on clinical data. To compare consistently, all the models were trained on the same training dataset and used the same training procedure.

Segmentation performance

Every test image was split into hundreds of 3D patches with overlaps and was processed by our segmentation model. The patches that contain IAs are called positive patches and those containing no aneurysm are called negative patches. To focus on the positive patches both in training and evaluation of the model, positive patches were duplicated manually to match approximately the same number as negative patches. Next, we evaluated the segmentation performance of our model compared with other models. The evaluation results are shown in Table 2. Our GLIA-Net increased the segmentation performance by a large margin in almost all metrics on the internal test set and two external datasets. Note that all the tested models have a better performance on the external test set A than on the internal test set, which may be because that the IAs in the external test set A are easier to identify (with bigger average size, as shown in Figure 3D). On external test set B, the performance of our model is only slightly lower than that on the internal test set, while other models drop a lot, which verifies the generalization ability of our model. The precision-recall curve and the ROC curve of our model are shown in Figure S1.

Detection performance

One of the aims of our model is to help clinicians find all the IAs hidden in the original CTA images. So, the performance of IA detection is also important. Unlike voxel-wise segmentation, target-wise detection is designed to point out where IAs exist, in which the IA shapes and sizes do not matter. Here, we set a standard to identify correct detection. First, we find out all the lesion regions in the segmentation map generated by the model and the ground truth label map. Then, if the center distance of any two lesion regions in the segmentation map and the ground truth map is smaller than the summation of their radiuses, we marked them matched. Then the matched and unmatched lesion regions can be used to calculate the following detection metrics, such as precision and recall.

The detection performance of our model and the comparison with other methods are shown in Table 3. Because U-Net and HeadXNet treat all patches from different parts of the CTA image in the same way, they may generate many false-positive predictions everywhere in the global image. Thanks to the global positioning network, our GLIA-Net not only reduced the false-positives a lot but also detected more IAs successfully. Specifically, our model detected 103 (98.5–108.3, 95% CI) IAs out of 126 in the internal test set, with only 4.38 (2.91–5.85, 95% CI)

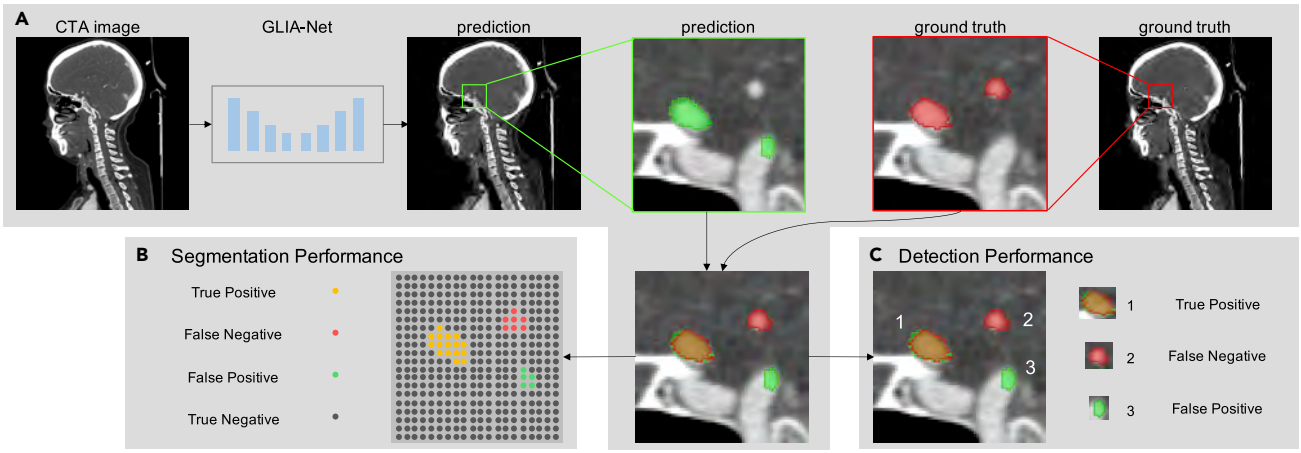


Figure 4. Evaluation procedures for voxel-wise segmentation performance and target-wise detection performance

(A) Generation of the prediction map using our GLIA-Net and the ground truth map on which we perform the evaluation.

(B) Segmentation performance is calculated on every voxel in the image, which measures the shape and edge of predicted IAs.

(C) Detection performance is calculated on every IA in the prediction and ground truth maps, which measures the ability of our model to identify IAs in CTA images.

false-positive predictions per case. The performance on the external test sets are slightly lower, but still comparable with that on the internal test set.

Results analysis

Besides the quantitative evaluations stated above and the ablation study of different parts of our GLIA-Net in Table S1 and Figure S1, we also show the visualization of four CTA images and the corresponding segmentation results predicted by our GLIA-Net and other methods in Figure 5. These examples were selected randomly from the internal test set. As shown in the figures, our model can identify IAs accurately even if they are ridiculously small compared with the whole CTA scan. With the help of our global positioning network, our GLIA-Net can reduce lots of false-negative predictions at low-risk regions where other models processing directly on local images often fail. Moreover, because the training is less affected by structures at low-risk regions, the true-positive predictions of our model are closer to the ground truth than other methods.

The global positioning network in our model can help the local segmentation network to learn global distribution information for the occurrence of IAs. This is important for current patch-based segmentation models on this task, which can only process one small local image clipped from the global image at a time and will lose the global structure information of the human body. To show the effectiveness of our global positioning network, we provide a visualization of its output global risk distribution map in Figure 6. The risk distribution map is built as follows. First, we collected the output probability from the final layer of our global positioning network for a small patch, which can represent the risk probability for this patch. Second, we collected the risk probabilities of all the tiled patches that can cover the whole global image. Finally, the value in each voxel in the probability distribution map was averaged by values of all the patches that contain this voxel. Because the small patches are tiled from the global image using the sliding-widow approach, we set the overlap of the sliding window to about 84% of the patch size to increase the resolution of the risk probability distribution map.

Red regions in Figure 6 indicate higher possibilities to contain IAs and blue regions are the relatively safe areas. However, note that the IAs are generally too tiny to spot in the down-sampled global images, which means the heatmap does not point out exactly the location of IAs, but a risk distribution in the whole scan. Most parts in the image are predicted low-risk areas of IAs inferred by our model, while most IAs occurred in high-risk regions for both the ground truth and the prediction of our model. The results show that the global positioning network can perceive human structural knowledge in the global image and extract useful information that is not opposed to the local segmentation network. This is important because similar structures to the IAs in other parts of the body can affect the training and testing procedure a lot if only performed on the local image.

Clinical experiment

Besides the performance on voxel-wise segmentation and target-wise detection, we also want to explore the clinical utility of our model and find out how the model can help clinicians. We performed a clinical experiment to compare the accuracy and time spent by radiologists to segment IAs with and without the assistance of our model. Without the assistance, the clinicians needed to perform the judgment by themselves from the raw CTA images. With the assistance, both the raw CTA images and the prediction of our model were available for the clinicians to diagnose, in which they can identify whether the results generated by our model are correct or not.

The design of our clinical experiment is shown in Figure 7. To simulate the real environment in clinical usage, we collected another 24 CTA scans from daily examinations in one of the internal institutions (Guizhou Provincial People's Hospital), of which 17 contain at least one IA and the others contain none. These CTA images were performed using our segmentation model in advance. The time spent only consists of the diagnosis and segmentation time by the clinicians without data-loading time. There were six resident radiologists and six attending radiologists from three institutions (Guizhou Provincial People's Hospital, The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine,

Table 2. Voxel-wise segmentation performance

Dataset	Model	Precision↑ (95% CI)	Recall↑ (95% CI)	DSC↑ (95% CI)	95% HD↓ (95% CI)	AUC↑ (95% CI)	AP↑ (95% CI)
Internal test	U-Net	14.0 (11.9–16.2)	71.3 (63.9–78.7)	23.2 (20.5–25.9)	19.6 (17.9–21.3)	98.8 (98.6–99.0)	17.5 (14.6–20.4)
	HeadXNet	16.2 (13.1–19.2)	55.6 (33.0–78.2)	23.2 (20.6–25.9)	15.9 (14.6–17.1)	98.2 (97.1–99.2)	25.0 (12.0–38.0)
	GLIA-Net (ours)	48.8 (44.5–53.0)	72.9 (66.9–78.9)	57.9 (56.4–59.5)	9.07 (7.84–10.3)	98.2 (97.6–98.8)	61.9 (59.4–64.4)
External test A	U-Net	23.9 (20.7–27.1)	71.0 (59.3–82.8)	35.3 (31.8–38.9)	19.5 (17.7–21.3)	97.9 (97.7–98.2)	30.2 (21.9–38.4)
	HeadXNet	27.1 (20.6–33.6)	52.7 (29.2–76.2)	32.4 (25.5–39.3)	15.6 (14.2–17.0)	96.2 (92.8–99.6)	32.1 (18.6–45.6)
	GLIA-Net (ours)	71.2 (65.2–77.3)	83.9 (82.2–85.7)	76.8 (73.7–79.9)	8.28 (7.05–9.52)	99.0 (98.8–99.1)	80.5 (78.6–82.3)
External test B	U-Net	6.54 (5.65–7.44)	43.6 (37.9–49.3)	11.3 (9.94–12.7)	21.0 (19.7–22.3)	86.9 (83.0–90.7)	6.86 (5.22–8.49)
	HeadXNet	14.8 (10.2–19.4)	32.6 (22.3–43.0)	18.1 (15.7–20.6)	15.9 (14.5–17.2)	86.1 (81.3–91.0)	14.3 (8.69–19.9)
	GLIA-Net (ours)	59.8 (54.7–64.8)	57.4 (43.0–71.8)	57.2 (50.5–64.0)	8.78 (7.79–9.76)	98.2 (97.2–99.2)	55.8 (44.0–67.6)

Values are given in units of % except for 95% HD, which is given in mm.

and Renhuai City People Hospital) participating in this experiment. Only the raw CTA images and the prediction maps were accessible to these radiologists during the experiment, which means they could not refer to the patient identifications, clinical reports, treatment histories, or follow-up examinations. To avoid the bias of different clinicians and different study cases, we used a crossover experiment design. We split the radiologists into two groups, R1 and R2, both of which had six radiologists. The study cases were also grouped into S1 and S2 equally. The radiologists in R1 diagnosed S1 without the assistance of our model and diagnosed S2 with the assistance. On the contrary, the radiologists in R2 diagnosed S1 with the assistance and diagnosed S2 without it.

We developed a CTA viewing and annotation tool by ourselves to assist radiologists in the IA diagnosis procedure (see Figure S6), which is also used in the clinical experiment. The evaluation result is shown in Table 4. We show the result in three aspects: (1) case-wise, which is to indicate the diagnostic ability of the radiologist; (2) target-wise, which shows the performance to detect IA targets; (3) voxel-wise, which is to evaluate the segmentation result that radiologists know about the shape and size of the aneurysms. Most of the diagnosing metrics with the model's assistance are better than without it, especially on the time spent, target-wise recall rate, and case-wise sensitivity. Our model is more helpful for resident radiologists than attending radiologists. The diagnosing time and target-wise recall rate of radiologists have a statistically significant difference before and after model assistance (p value 0.02 and <0.01). We also give an evaluation for radiologists from different institutions without and with the assistance of our model (see Table S2). Although there are diagnosing differences among different institutions, which may be due to different diagnosis abilities of different institutions, all radiologists can benefit from our model consistently, which also indicates that our model can help radiologists with different diagnosing performance in clinical practice.

The clinical experiment verifies our model's clinical utility. Our model helps clinicians increase their diagnosis performance in almost all metrics and saves diagnosis time. We should also note that the radiologists in this clinical experiment only need to identify IAs, but the most common situation for radiologists is to examine a variety of diseases, in which case they may need more time, and the assistance of the model should also be very helpful.

DISCUSSION

In this work, we collected a large IA segmentation dataset of CTA images with pixel-wise segmentation labels. Our internal dataset contains 1,338 CTA examinations consisting of 699,266 image slices and 1,489 IAs overall. Then we proposed a fully automatic segmentation technique that contains a global positioning network to provide global risk probability information and a patch-based local segmentation network to generate voxel-wise predictions. Our model can be directly used in different clinics and under different scan settings with no pre- or postprocessing procedure. The result shows that our GLIA-Net can identify over 80% of IA targets with only about four false-positives per case on the internal test set.

The improvement of our model can be explained by its global focus design. Different from most of the medical image segmentation networks directly transformed from those for natural images, our model makes use of the statistics of medical images as much as possible. Our design has several advantages in IA segmentation. First, the global positioning network combining with the patch-based local segmentation network leverages the full resolution of the CTA images without down-sampling the images to fit the memory limit. Otherwise, small targets like early-stage IAs or those tiny ones hiding in the corners of the brain may be lost. Second, such a global mechanism can provide local patches with a global location feature, which is important for many localization tasks such as identifying IAs in the brain, because such lesions have a strong tendency to locate in specific regions. Patch-based segmentation methods can work well for some segmentation tasks with subtle relation between the lesion region and its global position, such as segmentation on tumor pathology whole-slide images. However, IAs are related to the vascular network distributed in the brain, which makes the global position information important. Third, it is not rare for the brain CTA images to contain necks or even part of the lungs in real-world clinical applications. The global positioning network can be treated as an additional brain detection network, but it can give more information than just an object detection model and does not need additional annotations. Besides, a pyramid-weighted loss strategy was used in training, which can reduce the impact of the variability and uncertainty of expert labelers. The evaluation result shows that our

Table 3. Target-wise detection performance

Dataset	Model	TPs↑ (95% CI)	FPs↓ (95% CI)	FNs↓ (95% CI)	Recall↑ (95% CI)	FPs per case↓ (95% CI)
Internal test	U-Net	92.4 (89.0–95.8)	4.68k (3.72k–5.64k)	33.6 (30.2–37.0)	73.3 (70.6–76.0)	30.8 (24.5–37.1)
	HeadXNet	69.2 (45.1–93.3)	2.41k (1.21k–3.62k)	56.8 (32.7–80.1)	54.9 (35.8–74.1)	15.9 (7.96–23.8)
	GLIA-Net (ours)	103 (98.5–108)	666 (443–889)	22.6 (17.7–27.5)	82.1 (78.2–86.0)	4.38 (2.91–5.85)
External test A	U-Net	34.0 (30.6–37.4)	1.41k (1.07k–1.75k)	16.0 (12.6–19.4)	68.0 (61.2–74.8)	19.8 (15.0–24.7)
	HeadXNet	24.2 (15.9–32.5)	670 (335–1.01k)	25.8 (17.5–34.1)	48.4 (31.8–65.0)	9.44 (4.71–14.2)
	GLIA-Net (ours)	36.0 (33.9–38.1)	193 (127–259)	14.0 (11.9–16.1)	72.0 (67.7–76.3)	2.72 (1.79–3.65)
External test B	U-Net	31.8 (29.4–34.2)	3.08k (21.3k–4.02k)	19.2 (16.8–21.6)	62.4 (57.7–67.0)	45.9 (31.8–60.0)
	HeadXNet	24.8 (18.5–31.1)	1.64k (737–2.55k)	26.2 (19.9–32.5)	48.6 (36.3–60.9)	24.5 (11.0–38.0)
	GLIA-Net (ours)	36.6 (33.3–39.9)	296 (222–371)	14.4 (11.1–17.7)	71.8 (65.4–78.2)	4.42 (3.31–5.54)

The total number of true positive (TP), false-positive (FP), and false-negative (FN) predicted IAs are shown. The target-wise recall rate in the unit of % and the number of false-positive IAs per case are also given.

technique is more capable in this IA segmentation task than other methods.

In clinical practice, the size of IAs also matters in the diagnosing procedure. Small IAs (such as smaller than 3mm) are relatively hard to identify, while large IAs are hardly missed. In our experiment, we also quantify this influence of IA sizes (see Table 5 and Figure S2). The experiment shows that the recall rate of IAs larger than 7 mm is about 20% higher than that of IAs smaller than 3 mm for our GLIA-Net. Although U-Net can achieve a compatible recall rate for targets smaller than 3 mm and larger than 7 mm, it generates over 30 false-positive predictions per case, which is not applicable for clinical usage. Our GLIA-Net can identify most of the aneurysms of different sizes while keeping the number of false-positive targets much smaller.

Although our GLIA-Net achieved promising results in IA segmentation and could assist radiologists and clinicians in their diagnosis, it also has some limitations. First, during the COVID-19 pandemic, performing a clinical study with more radiologists and larger clinical cases is much more difficult than usual. However, a continuous online validation enrolling more radiologists and more cases with the assistance of our model in clinical practice would better demonstrate the validity of our model in the future. Second, our model was only trained on CTA images. If other modalities, such as MRA images and DSA (the current gold standard for IA diagnosis), are adopted, the usage scope can be broader, and the accuracy may be higher.

The study of IA is important for the health of the public. In the future, beyond the identification and segmentation, predicting the rupture of IAs based on the knowledge of the location, size, and shape of IAs will be worth exploring. Therefore, GLIA-Net can not only assist clinicians in the diagnosis of IAs but can also encourage more implementations of artificial intelligence in healthcare.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and materials should be directed to and will be fulfilled by the Lead contact, Feng Xu (feng-xu@tsinghua.edu.cn).

Materials availability

This study did not generate any materials.

Data and code availability

Original data have been deposited to our data server: http://39.98.209.108/seafile/d/34d94c4bc44a42fdb33d/. The internal and external datasets are only available for non-commercial purposes.

The GLIA-Net, together with the training and evaluation code generated during this study, are available at Github: https://github.com/MeteorsHub/GLIA-Net. The annotation tool used in the clinical study is also available at Github: https://github.com/MeteorsHub/MedLabelMe.

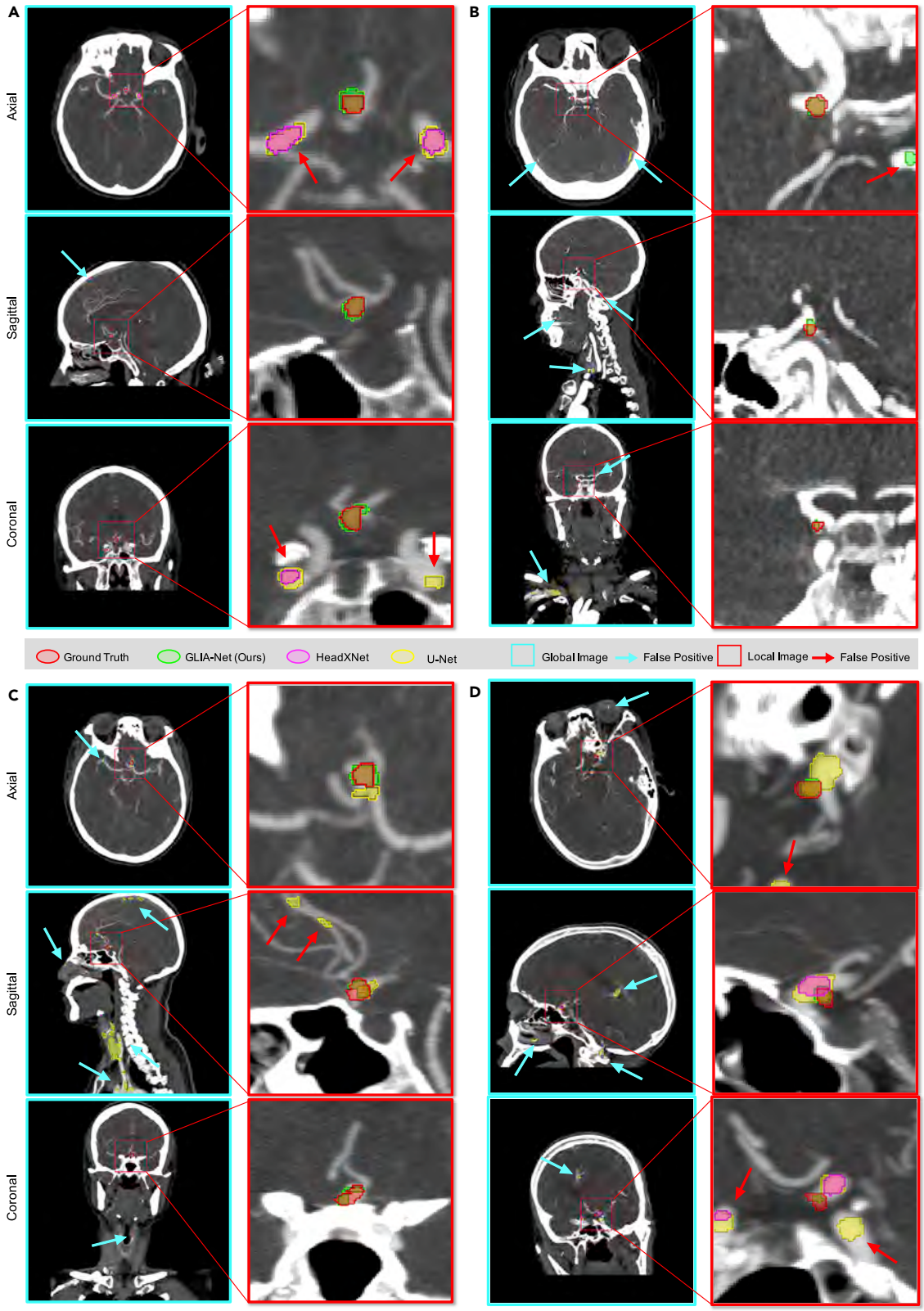
Methods

Deep learning usually uses artificial neural networks with millions and billions of parameters to fit a mathematical function to predict or describe a specific problem. The network needs a lot of training data to modify its parameters and generate outputs closer and closer to the real outputs using the gradient descent method. In our case, we used a deep neural network to segment every IA with different sizes, shapes, and locations in the large CTA image scans.

The whole pipeline of our method is shown in Figure 1. We use GLIA-Net (global positioning network) to segment IAs from CTA image scans. It takes the whole 3D CTA scan, called global image as input, which has a variant number of slices along the depth axis. GLIA-Net tiles many sub-images, or patches, from the global image with an overlap size to process, which are called the local images. GLIA-Net consists of two dataflow pathways: the global positioning network and the local segmentation network. The global positioning network is proposed to estimate the global probability distribution of IAs in the global image. The local segmentation network is focused on local images that are tiled from the global image and gives voxel-wise segmentation of the IA. During the training period, a global binary loss was added to the global positioning network and a local voxel-wise segmentation loss to the local segmentation network. Besides, we propose a pyramid-weighted loss that takes the variability and uncertainty of expert labelers into account in the loss design of the local segmentation network. The data preprocessing and network details are elaborated below and in the Supplemental experimental procedures.

Global positioning network

Because our GLIA-Net processes every local image patch one by one to segment IA, the global location information for the current patch will be lost if no additional global feature is adopted. Thus, a global positioning network is proposed to extract a global positioning feature and feed it to the local segmentation network. First, several 3D convolution blocks are applied to the global image input and extract a global feature map. This global feature map is expected to have global distribution information about how IAs locate in the brain and neighboring body parts. Then, a region-of-interest pooling layer[41] is added to pool the feature map from the bounding box corresponding to the current local image location. The extracted global feature map for the current patch location is then reshaped to a fixed size by an adaptive maximum-pooling layer, and processed by another few 3D convolution blocks to generate the global location feature, which will be sent to the local segmentation network to assist in segmenting fine-scale targets.



(legend on next page)

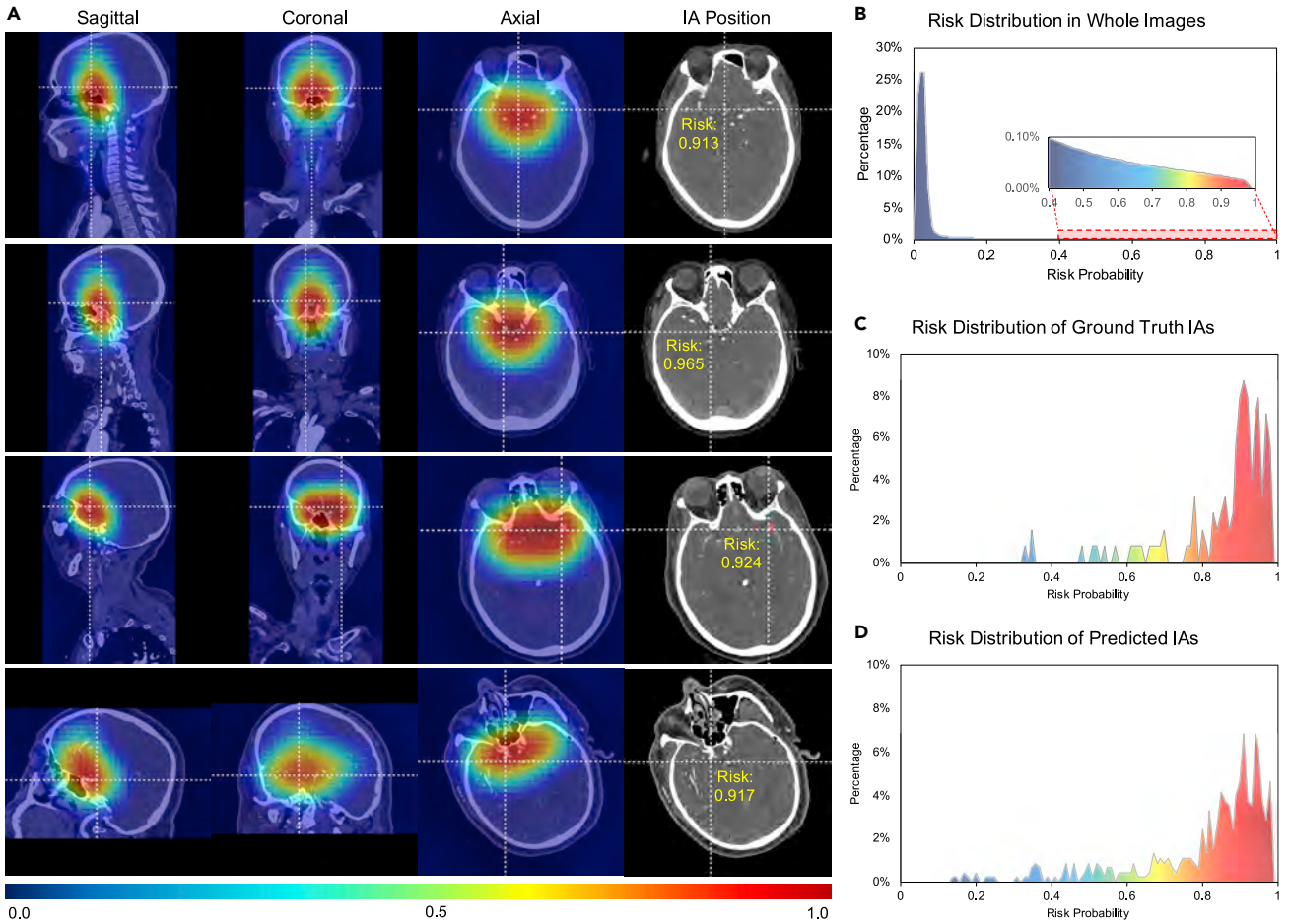


Figure 6. Global probability distribution generalized by the global positioning network

(A) Four examples in four rows are randomly selected from the internal test set. Generated by the global positioning network, the probability distribution maps are shown in false color and displayed on top of the original CTA image. The first three columns are the sagittal, coronal, and axial perspectives. The final column shows the actual position of the IAs using crosshairs, together with the risk probability value at the target center.

(B–D) (B) The risk value distribution in whole CTA images, which summarizes all the voxels of cases in the internal test set. The percentages of high-risk regions are quite small in the whole scans. The risk value distributions of the center of (C) ground truth IAs and (D) predicted IAs by our GLIA-Net are summarized on all the IAs. Most IAs occurred in high-risk regions and our model estimates the IA risk distribution successfully.

In order to guide the global positioning network with an intermediate objective, we add a global feature classification cross-entropy loss. If there are any IA voxels in the current processing local patch, the global label is positive, otherwise negative. A 3D convolution layer, a global maximum-pooling layer, and a fully connected layer are added to the global position feature map, and a softmax cross-entropy loss is computed as the global positioning loss:

$$\mathcal{L}_{Global} = -(z \ln \hat{z} + (1-z)\ln(1-\hat{z}))$$

where z and $\hat{z}$ are the ground truth label and the softmax probability prediction of the global positioning network output.

Local segmentation network

Our objective is the 3D segmentation of IA on CTA scans, which is a difficult segmentation task because the target region is quite small. So, our local segmentation network is applied to the local images under the original image resolution to avoid missing any small target. We modify U-Net[18] to serve as our local segmentation network, which is an encoder-decoder structure. All the convolution and pooling layers are transformed into a 3D version, and we use residual network blocks[42] to avoid gradient vanishing in our deep model. The encoder architecture consists of several levels of layers to extract image features in different scales. As for the decoder, we adopt transpose convolution layers to recover segmentation results step by step. Skip connections are

Figure 5. Qualitative comparison with other models

Segmentation results of four CTA cases, (A), (B), (C), and (D), randomly selected from the internal test set. There are three perspectives (axial, sagittal, and coronal) for each 3D CTA case for better visualization. The CTA images are shown using 11-layer MIP (maximum intensity projection). The window width and window level of the CTA images are 600 and 200 Hounsfield Units. Because of the global positioning network, our GLIA-Net has much fewer false-positive predictions at low-risk regions where other methods often generate positive, including in-head (e.g., sagittal in [A] and axial in [B]) and out-of-head (e.g., coronal in [B] and sagittal in [C]) regions. Besides, the segmentation results of our model are also closer to the annotation compared with other models (e.g., local images in [C] and [D]). Better viewed in a high-resolution image.

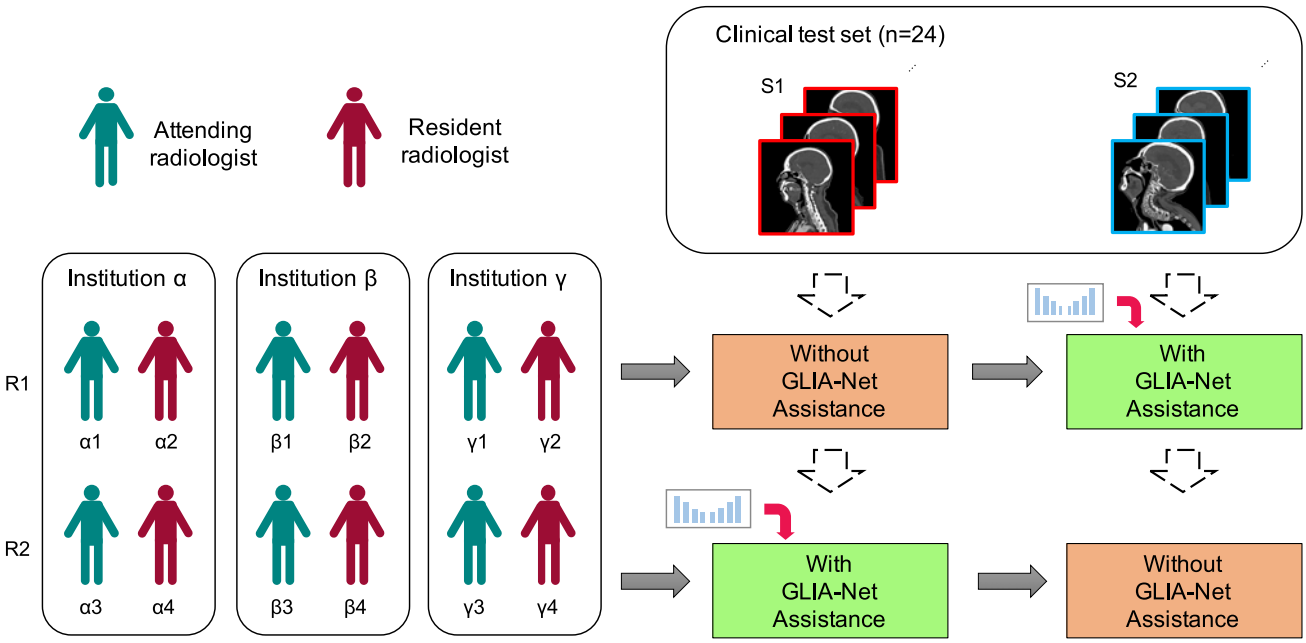


Figure 7. Clinical study design

We use a crossover clinical study design in which six attending and six resident radiologists from three different institutions were split into R1 and R2, while 24 CTA cases were split into S1 and S2. The clinical test dataset was collected from daily examinations in one of the internal institutions, which is an independent validation dataset.

used to link each layer level between the encoder and the decoder. The location feature extracted from the global positioning network is element-wise multiplied to all the skip connections before adaptive pooling layers to unify their shapes.

Because of the small target size compared with the input image, we use exponential logarithmic loss[43] that combines cross-entropy loss and dice loss for our local segmentation network:

$$\mathcal{L}_{Local} = \omega_{Dice}\mathcal{L}_{Dice} + \omega_{Cross}\mathcal{L}_{Cross}$$

where ω_{Dice} and ω_{Cross} are the weights for the dice loss and cross-entropy loss, respectively. The dice loss is used to minimize the shape difference of the segmentation output and the ground truth target:

$$\mathcal{L}_{Dice} = E\left(-\ln\frac{2y\hat{y}+\varepsilon}{y+\hat{y}+\varepsilon}\right)^{\gamma_{Dice}}$$

where $E(*)$ is the function to compute the mean value for each voxel position. y and $\hat{y}$ are the ground truth label and the probability prediction of the local

Table 4. Diagnosing difference with and without the assistance of our model in our clinical study

			Voxel-wise	Target-wise		Case-wise		
		Time↓	DSC↑ (95% CI)	Precision↑ (95% CI)	Recall↑ (95% CI)	Specificity↑ (95% CI)	Sensitivity↑ (95% CI)	ACC↑ (95% CI)
Model		25.8 (24.9–26.7)	30.1 (25.3–34.8)	19.3 (14.2–24.4)	85.8 (80.0–91.5)	38.9 (26.1–51.7)	96.2 (92.2–100)	79.2 (75.6–82.8)
Attending radiologist	without assist	132 (118–145)	57.9 (46.2–69.6)	91.7 (82.0–100)	68.8 (52.2–85.3)	100 (100–100)	80.6 (67.8–93.3)	86.1 (76.9–95.3)
	with assist	121 (107–135)	67.6 (58.8–76.4)	91.7 (82.0–100)	88.2 (79.1–97.3)	90.3 (79.1–100)	88.7 (79.8–97.6)	88.9 (79.7–98.1)
Resident radiologist	without assist	162 (150–175)	48.2 (29.0–67.4)	81.6 (68.3–94.9)	66.0 (48.1–83.9)	83.3 (63.0–100)	82.4 (68.5–96.3)	83.3 (72.4–94.2)
	with assist	143 (131–156)	54.4 (40.3–68.6)	88.4 (77.4–99.4)	84.7 (77.6–91.8)	100 (100–100)	84.9 (78.1–91.6)	88.9 (83.9–93.9)
All	without assist	147 (137–156)	53.1 (41.5–64.6)	86.6 (77.9–95.3)	67.4 (55.2–79.6)	91.7 (80.4–100)	81.5 (72.1–90.9)	84.7 (77.6–91.9)
	with assist	132 (123–142)	61.0 (51.9–70.1)	90.0 (82.6–97.4)	86.5 (80.6–92.3)	95.1 (88.9–100)	86.8 (81.1–92.4)	88.9 (83.7–94.1)
	p value	0.02	0.16	0.29	<0.01	0.31	0.19	0.19

Time is given in seconds. p value is computed on 12 radiologists. Other metrics are given in %. ACC, accuracy.

Table 5. Detection performance for aneurysms with different sizes on the internal test dataset

Aneurysm Size (mm)	Metric	U-Net	HeadXNet	GLIA-Net
<3	recall↑	74.6 (70.0–79.2)	53.5 (33.6–73.5)	70.3 (64.1–76.5)
	FPs per case↓	35.9 (28.4–43.4)	15.6 (7.73–23.4)	5.22 (3.52–6.92)
3–7	recall↑	70.7 (67.3–73.8)	54.8 (36.0–73.7)	82.3 (77.8–86.7)
	FPs per case↓	36.5 (28.7–44.2)	18.2 (9.47–27.0)	5.05 (3.40–6.71)
>7	recall↑	91.8 (89.2–94.3)	67.1 (49.0–85.1)	90.6 (88.1–93.1)
	FPs per case↓	32.1 (24.6–39.7)	15.2 (8.31–22.2)	3.52 (2.32–4.72)

FPs per case is the number of false-positive targets predicted per case. The unit of the recall rate is %.

segmentation network output. γ_{Dice} is the parameter to control the nonlinearities of the loss function. ε is used as a smoothing factor.

Besides, we propose a pyramid weighting strategy for the cross-entropy loss. As mentioned above, the IAs are exceedingly small in the brain. This causes a problem for radiologists in the dataset annotation procedure: they may mark a precise IA position but with a relatively vague segmentation edge considering the few voxels in that region. In the meantime, human experts' annotation has high inter-observer variability,[32] which means they may not share the same labeling standard on the lesion edges. To solve the label uncertainty problem for small segmentation regions (see details in Figures S3 and S5), we adopt a pyramid-weighted cross-entropy loss in our IA segmentation loss. The computation detail of the pyramid weight is shown in Figure S4. The weights of the loss for voxels near the center of IAs are high and those to the IA edges are low. In consideration of not affecting the size of the segmentation results of large targets, where the boundary is not the crucial component, we only compute pyramid weights for small targets less than 400 voxels, and the larger ones are set with a fixed weight. Then the loss function for cross-entropy is:

$$\mathcal{L}_{Cross} = E(\omega_p(-(y\ln\hat{y}+(1-y)\ln(1-\hat{y})))^{\gamma_{Cross}})$$

where ω_p is the pyramid weight. To build the pyramid weights, we use minimum pooling with a 3 × 3 kernel to erode the label map step by step until the center of the region, and then sum up all the intermediate label maps. This training technique leads the network to focus on positive region locations with less attention to the lesion edges.

At last, we train our end-to-end model using a combined loss:

$$\mathcal{L}_{Total} = \omega_{Global}\mathcal{L}_{Global} + \omega_{Local}\mathcal{L}_{Local}$$

where ω_{Global} and ω_{Local} are the loss weights for global loss and local loss respectively.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethical approval was obtained from the ethics committee at each institution. Informed consent was obtained from participants at each institution for ethical approval. In this study, a de-identification method was introduced to protect patient privacy by removing sensitive information (medical record number, account number, patient name, birthday, contact, address, etc.) manually from medical data.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100197.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2018YFA0704000), the National Natural Science Foundation of China (61822111, 81960314, 62071271), the National Postdoctoral Program for Innovative Talent (BX20190173), the Beijing Natural Science Foundation (JQ19015), the Science and Technology Foundation of Guizhou Province (QKHZC[2019]2810, QKHJC[2016]1096, QKHPTRC[2019]5803, and QKHPTRC[2017]5724), and the Guizhou Science and Technology Department Key Lab Project (QKF[2017]25).

We appreciate the institutions for providing the internal and external datasets in this work, which include Guizhou Provincial People's Hospital, Affiliated Hospital of Zunyi Medical University, Tongren Municipal People's Hospital, Xingyi Municipal People's Hospital, The Second People's Hospital of Guiyang, The First People's Hospital of Zunyi, People's Hospital of Anshun City, and Zhijin People's Hospital. We also appreciate the radiologists enrolled in the clinical experiment from Guizhou Provincial People's Hospital, The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, and Renhuai City People Hospital.

AUTHOR CONTRIBUTIONS

Conceptualization, F. X., R.P.W., and Q.H.D.; Methodology, Z.-H.B. and F.X.; Software, Z.-H.B.; Validation, Z.-H.B., C.T., and W.C.L.; Formal Analysis, Z.-H.B. and H.Q.; Investigation, Z.-H.B., H.Q., and C.T.; Resources, C.T., W.C.L., T.T.L., D.X.L., D.L., X.C.Z., L.L.M., T.L.S., B.W., C.H., L.L., C.J., and Q.P.G.; Data Curation, C.T. and W.C.L.; Writing – Original Draft, Z.-H.B. and C.T.; Writing – Review & Editing, H.Q. and F.X.; Visualization, Z.-H.B. and C.T.; Supervision, H.Q., Y.C.G., and F.X.; Project Administration, J.-H.Y., T.J.Z., R.P.W., and Q.H.D.; Funding Acquisition, H.Q., F.X., R.P.W., and Q.H.D.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 16, 2020
Revised: October 1, 2020
Accepted: December 29, 2020
Published: January 22, 2021

REFERENCES

1. Jayaraman, M.V., Mayo-Smith, W.W., Tung, G.A., Haas, R.A., Rogg, J.M., Mehta, N.R., and Doberstein, C.E. (2004). Detection of intracranial aneurysms: multi–detector row CT angiography compared with DSA. Radiology *230*, 510–518.
2. Eskey, C.J., Meyers, P.M., Nguyen, T.N., Ansari, S.A., Jayaraman, M., McDougall, C.G., DeMarco, J.K., Gray, W.A., Hess, D.C., and Higashida, R.T. (2018). Indications for the performance of intracranial endovascular neurointerventional procedures: a scientific statement from the American Heart Association. Circulation *137*, e661–e689.
3. Brown, R.D., Jr., and Broderick, J.P. (2014). Unruptured intracranial aneurysms: epidemiology, natural history, management options, and familial screening. Lancet Neurol. *13*, 393–404.
4. Lawton, M.T., and Vates, G.E. (2017). Subarachnoid hemorrhage. N. Engl. J. Med. *377*, 257–266.

5. Hop, J.W., Rinkel, G.I.J., Algra, A., and van Gijn, J. (1997). Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review. Stroke *28*, 660–664.

6. Takagi, K., Tamura, A., Nakagomi, T., Nakayama, H., Gotoh, O., Kawai, K., Taneda, M., Yasui, N., Hadeishi, H., and Sano, K. (1999). How should a subarachnoid hemorrhage grading scale be determined? A combinatorial approach based solely on the Glasgow Coma Scale. J. Neurosurg. *90*, 680–687.

7. Ueda, D., Yamamoto, A., Nishimori, M., Shimono, T., Doishita, S., Shimazaki, A., Katayama, Y., Fukumoto, S., Choppin, A., Shimahara, Y., et al. (2019). Deep learning for MR angiography: automated detection of cerebral aneurysms. Radiology *290*, 187–194.

8. Park, A., Chute, C., Rajpurkar, P., Lou, J., Ball, R.L., Shpanskaya, K., Jabarkheel, R., Kim, L.H., McKenna, E., Tseng, J., et al. (2019). Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. JAMA Netw. Open *2*, e195600.

9. Nakao, T., Hanaoka, S., Nomura, Y., Sato, I., Nemoto, M., Miki, S., Maeda, E., Yoshikawa, T., Hayashi, N., and Abe, O. (2018). Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J. Magn. Reson. Imaging *47*, 948–953.

10. Dai, X., Huang, L., Qian, Y., Xia, S., Chong, W., Liu, J., Di Ieva, A., Hou, X., and Ou, C. (2020). Deep learning for automated cerebral aneurysm detection on computed tomography images. Int. J. Comput. Assist. Radiol. Surg. *15*, 715–723.

11. Westerlaan, H.E., Van Dijk, J., Jansen-van der Weide, M.C., de Groot, J.C., Groen, R.J., Mooij, J.J.A., and Oudkerk, M. (2011). Intracranial aneurysms in patients with subarachnoid hemorrhage: CT angiography as a primary examination tool for diagnosis—systematic review and meta-analysis. Radiology *258*, 134–145.

12. Ozpeynirci, Y., Braun, M., and Schmitz, B. (2019). CT angiography in occlusion assessment of intracranial aneurysms treated with the WEB device. J. Neuroimaging *29*, 481–486.

13. Anderson, G.B., Steinke, D.E., Petruk, K.C., Ashforth, R., and Findlay, J.M. (1999). Computed tomographic angiography versus digital subtraction angiography for the diagnosis and early treatment of ruptured intracranial aneurysms. Neurosurgery *45*, 1315–1322.

14. Kouskouras, C., Charitanti, A., Giavroglou, C., Foroglou, N., Selviaridis, P., Kontopoulos, V., and Dimitriadis, A. (2004). Intracranial aneurysms: evaluation using CTA and MRA. Correlation with DSA and intraoperative findings. Neuroradiology *46*, 842–850.

15. Ni, Q., Chen, G., Schoepf, U., Klitsie, M., De Cecco, C., Zhou, C., Luo, S., Lu, G., and Zhang, L. (2016). Cerebral CTA with low tube voltage and low contrast material volume for detection of intracranial aneurysms. Am. J. Neuroradiology *37*, 1774–1780.

16. Firouzian, A., Manniesing, R., Flach, Z.H., Risselada, R., van Kooten, F., Sturkenboom, M.C.J.M., van der Lugt, A., and Niessen, W.J. (2011). Intracranial aneurysm segmentation in 3D CT angiography: method and quantitative validation with and without prior noise filtering. Eur. J. Radiol. *79*, 299–304.

17. Zhai, X., Eslami, M., Hussein, E.S., Filali, M.S., Shalaby, S.T., Amira, A., Bensaali, F., Dakua, S., Abinahed, J., Al-Ansari, A., et al. (2018). Real-time automated image segmentation technique for cerebral aneurysm on reconfigurable system-on-chip. J. Comput. Sci. *27*, 35–45.

18. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. Med. Image Comput. Comput. Assist. Interv. *9351* (Pt Iii), 234–241.

19. Milletari, F., Navab, N., and Ahmadi, S.A. (2016). V-net: fully convolutional neural networks for volumetric medical image segmentation. Int. Conf. 3d Vis. 565–571.

20. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, eds. (Springer), pp. 424–432.

21. Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al. (2019). DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 1559–1572.

22. Emad, O., Yassine, I.A., and Fahmy, A.S. (2015). Automatic localization of the left ventricle in cardiac MRI images using deep learning. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (IEEE), pp. 683–686.

23. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., and Heng, P.A. (2018). H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging *37*, 2663–2674.

24. Zheng, Q., Delingette, H., Duchateau, N., and Ayache, N. (2018). 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Trans. Med. Imaging *37*, 2137–2148.

25. Kong, B., Zhan, Y., Shin, M., Denny, T., and Zhang, S. (2016). Recognizing end-diastole and end-systole frames via deep temporal regression network. In International Conference on Medical Image Computing and Computer-Assisted Intervention, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, eds. (Springer), pp. 264–272.

26. Wang, Z., Zou, N., Shen, D., and Ji, S. (2020). Non-local U-nets for biomedical image segmentation. In AAAI (AAAI Press), pp. 6315–6322.

27. Huang, Y., Dou, Q., Wang, Z., Liu, L., Wang, L., Chen, H., Heng, P.-A., and Xu, R. (2018). HL-FCN: hybrid loss guided FCN for colorectal cancer segmentation. In IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 195–198.

28. Huang, Y.-J., Dou, Q., Wang, Z.-X., Liu, L.-Z., Jin, Y., Li, C.-F., Wang, L., Chen, H., and Xu, R.-H. (2018). 3d RoI-aware U-Net for accurate and efficient colorectal tumor segmentation. arXiv, arXiv:1806.10342.

29. Anbeek, P., Vincken, K.L., Van Osch, M.J., Bisschops, R.H., and Van Der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR imaging. NeuroImage *21*, 1037–1044.

30. Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. Neuroinformatics *13*, 261–276.

31. Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I.W.M., Sanchez, C.I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., and Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. *7*, 5110.

32. Suinesiaputra, A., Bluemke, D.A., Cowan, B.R., Friedrich, M.G., Kramer, C.M., Kwong, R., Plein, S., Schulz-Menger, J., Westenberg, J.J., and Young, A.A. (2015). Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. J. Cardiovasc. Magn. Reson. *17*, 63.

33. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage *31*, 1116–1128.

34. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognit. Lett. *27*, 861–874.

35. Zhu, M. (2004). Recall, Precision and Average Precision, *2* (Department of Statistics and Actuarial Science, University of Waterloo), p. 30.

36. Raudaschl, P.F., Zaffino, P., Sharp, G.C., Spadea, M.F., Chen, A., Dawant, B.M., Albrecht, T., Gass, T., Langguth, C., and Lüthi, M. (2017). Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. Med. Phys. *44*, 2020–2036.

37. Huttenlocher, D.P., Klanderman, G.A., and Rucklidge, W.J. (1993). Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. *15*, 850–863.

38. Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 1492–1500.

39. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 7132–7141.

40. Chen, L.C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv, arXiv:1706.05587.

41. Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. *39*, 1137–1149.

42. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 770–778.

43. Wong, K.C., Moradi, M., Tang, H., and Syeda-Mahmood, T. (2018). 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In International Conference on Medical Image Computing and Computer-Assisted Intervention, A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds. (Springer), pp. 612–619.

PATTER, Volume *2*

Supplemental Information

Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network

Zi-Hao Bo, Hui Qiao, Chong Tian, Yuchen Guo, Wuchao Li, Tiantian Liang, Dongxue Li, Dan Liao, Xianchun Zeng, Leilei Mei, Tianliang Shi, Bo Wu, Chao Huang, Lu Liu, Can Jin, Qiping Guo, Jun-Hai Yong, Feng Xu, Tijiang Zhang, Rongpin Wang, and Qionghai Dai

Table S1. | Ablation study on the internal test dataset.

Metrics		GLIA-Net w/o global positioning network	GLIA-Net w/o pyramid weighted loss	GLIA-Net
Voxel-wise	Precision↑	**49.8 (38.6-60.9)**	60.2 (50.2-70.2)	48.8 (44.5-53.0)
	Recall↑	36.4 (23.8-49.0)	60.5 (56.2-64.7)	**72.9 (66.9-78.9)**
	DSC↑	40.3 (29.5-51.1)	**60.0 (53.0-67.0)**	57.9 (56.4-59.5)
	95%HD↓	9.80 (7.85-11.8)	**7.91 (6.57-9.25)**	9.07 (7.84-10.3)
	AUC↑	90.0 (81.5-98.4)	92.1 (90.8-93.5)	**98.2 (97.6-98.8)**
	AP↑	35.7 (23.3-48.2)	57.3 (50.5-64.1)	**61.9 (59.4-64.4)**
Target-wise	Recall↑	39.7 (17.6-61.8)	44.0 (39.4-48.6)	**82.1 (78.2-86.0)**
	FPs per case↓	**1.10 (0.36-1.84)**	3.72 (1.38-6.06)	4.38 (2.91-5.85)

FPs per case is the number of false positive predictions per case. 95%HD is given in mm. Other Values are given in units of %.

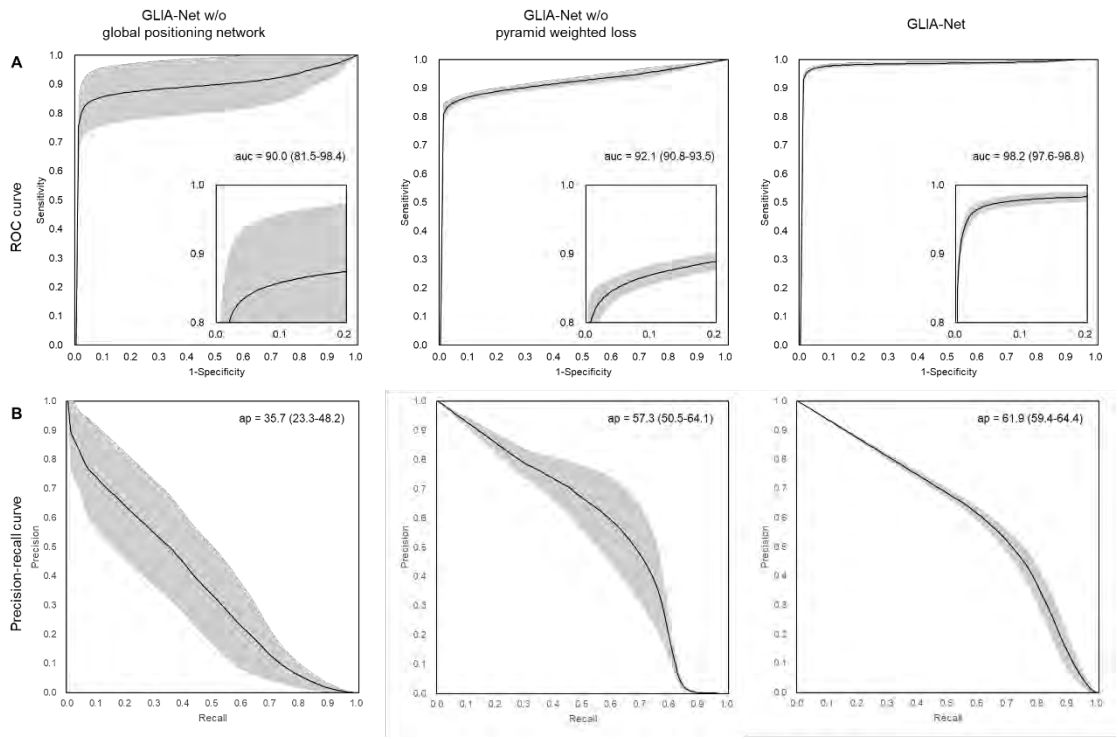


Figure S1. Segmentation performance of the ablation study on the internal test dataset. (A) ROC curve and (B) precision-recall curve of our GLIA-Net without global positioning network, our GLIA-Net without pyramid weighted loss, and our final GLIA-Net are shown. The AP and AUC values are given in “mean (95%CI)”. Most of the evaluation metrics get much worse without the support of our global positioning network.

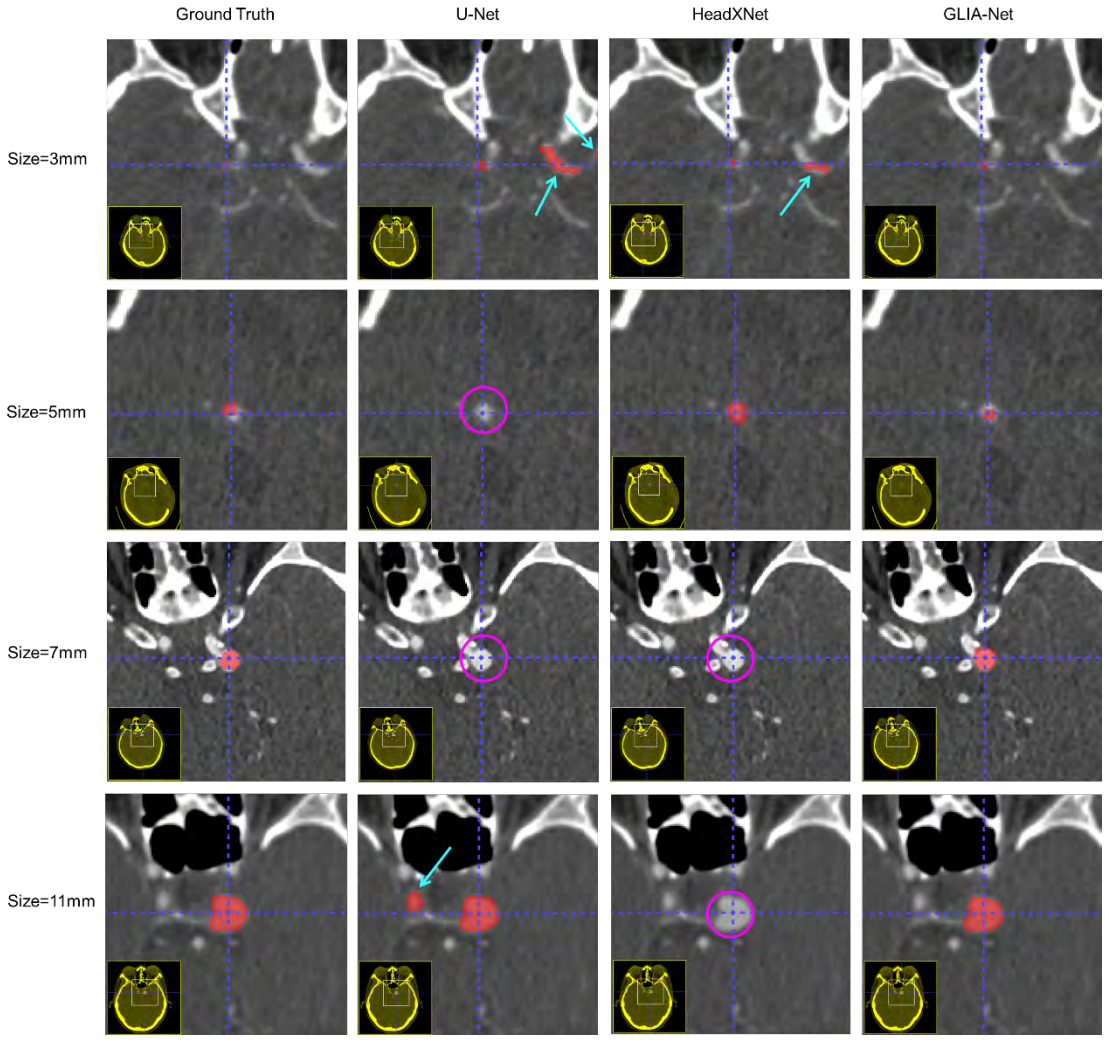


Figure S2. Segmentation results for 4 IAs of different sizes in the internal test dataset. The blue arrow points out the false positive predictions. The pink circle means the model fails to find the lesion area. The blue crosshair indicates the position of IAs.

Table S2. Clinical study performance of different institutions.

		Time↓	Voxel-wise	Target-wise		Case-wise		
			DSC↑ (95%CI)	Precision↑ (95%CI)	Recall↑ (95%CI)	Specificity↑ (95%CI)	Sensitivity↑ (95%CI)	ACC↑ (95%CI)
Institution α	Without Assist	147 (134-159)	70.9 (56.4-85.4)	90.9 (75.5-100)	83.3 (66.0-100)	83.3 (55.0-100)	**97.2 (92.5-100)**	93.8 (87.0-100)
	With assist	**123 (108-138)**	**74.8 (70.5-79.2)**	**93.2 (81.6-100)**	**95.8 (91.8-99.9)**	**100 (100-100)**	94.4 (89.0-99.9)	**95.8 (91.8-99.9)**
Institution β	Without assist	133 (115-151)	41.3 (23.8-58.8)	79.7 (73.2-86.3)	56.3 (49.5-63.0)	**100 (100-100)**	75.7 (62.8-88.6)	**83.3 (75.2-91.5)**
	With assist	**120 (105-134)**	**54.4 (44.7-64.1)**	**85.2 (74.2-96.3)**	**78.1 (70.3-86.0)**	85.4 (70.8-100)	**77.3 (69.9-84.7)**	79.2 (72.1-86.2)
Institution γ	Without assist	161 (144-179)	47.0 (32.4-61.5)	89.3 (71.1-100)	62.5 (37.8-87.2)	91.7 (77.5-100)	71.5 (55.9-87.2)	77.1 (62.5-91.7)
	With assist	**154 (136-172)**	**53.9 (34.9-72.9)**	**91.7 (77.5-100)**	**85.4 (76.5-94.3)**	**100 (100-100)**	**88.5 (80.8-96.3)**	**91.7 (85.9-97.4)**

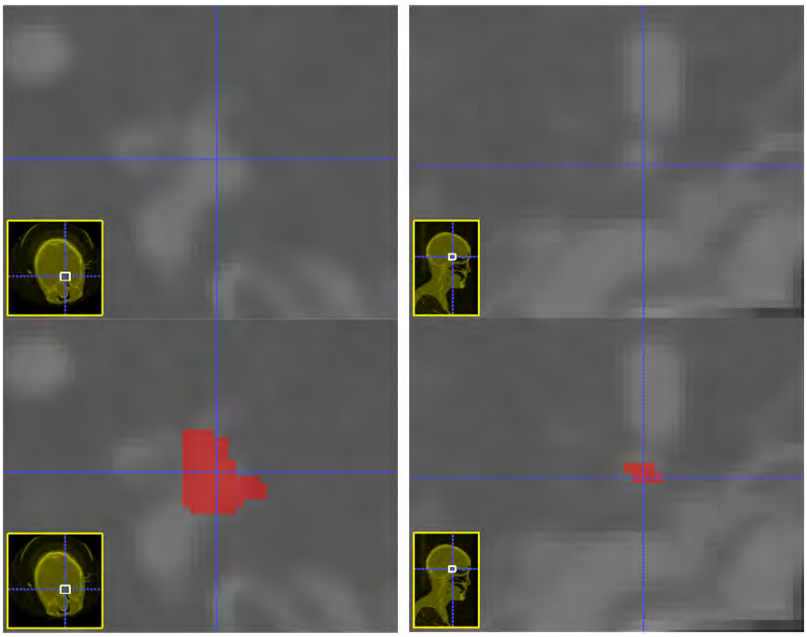


Figure S3. Annotation details for aneurysm segmentation. There are two label annotations achieved by radiologists in which the red mask is the annotation label. The boundary between aneurysms and brain tissue is very blurry, let alone that between aneurysms and their attached vascular, especially for the small case from the right figure. This is not a labeling error, but a result of the low resolution of CTA images and the definition of lesion regions. So we propose a pyramid weighted loss strategy to overcome this phenomenon.

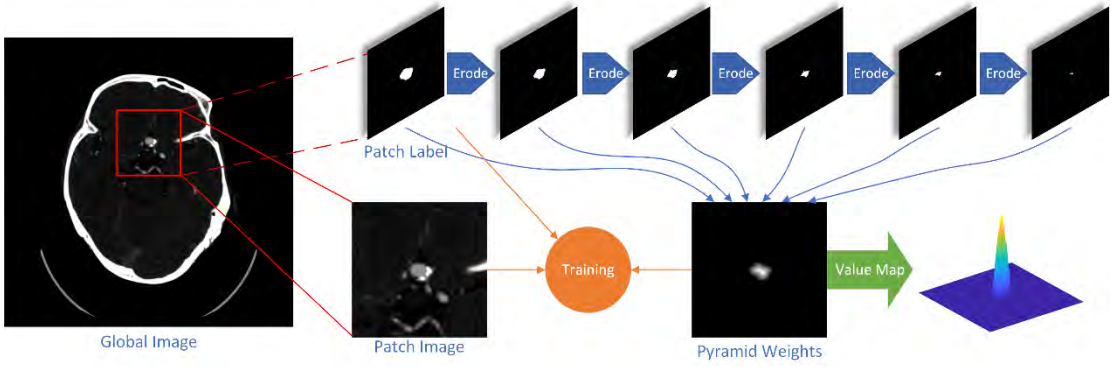


Figure S4. The pipeline of building the pyramid weights. The patch label map is eroded recursively and summed up to build the pyramid weights, which has high values in the target center, and low values on the edge. The highest and lowest weight values are fixed and the values in between are linearly scaled.

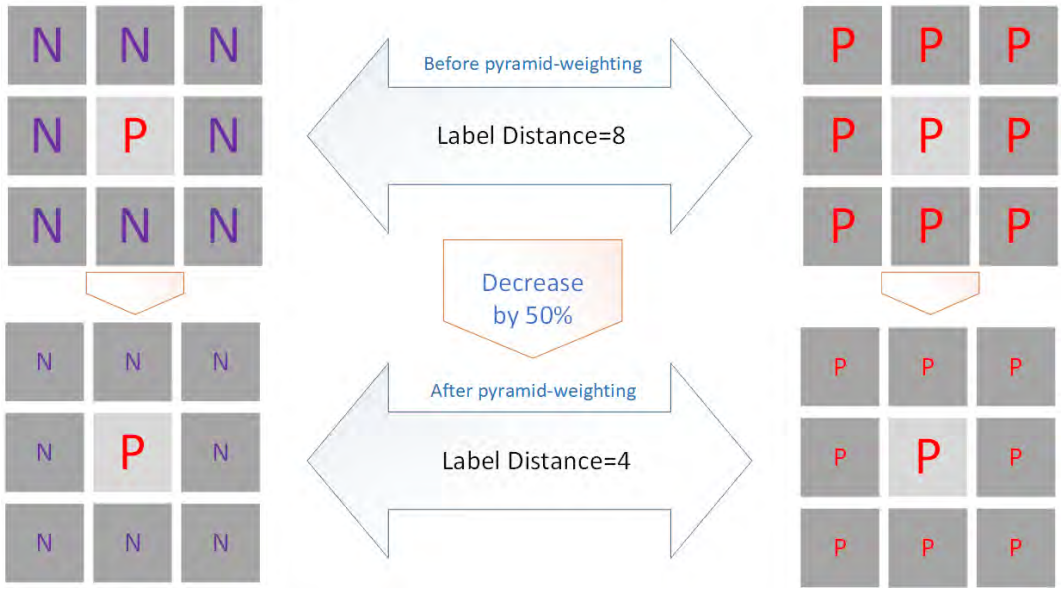


Figure S5. Label consistency before and after using the pyramid-weighted strategy. Each of the two image samples contains 9 pixels, which represents a small aneurysm with a ”certain area” in the center and “uncertain areas” on the edge. P and N represent positive and negative labels. Both the left and right samples may occur in the training set because of the different labeling standards. Before pyramid-weighting, all the pixels have the same loss weights of 1.0, leading a label distance between the two images to 8. After pyramid weighting, the weights of the neighboring pixels are decreased to 0.5, which decreases the label distance by 50%. With a low label consistency, the model may think these training samples to be noise and will not learn anything from them. This weighting strategy can increase the labeling consistency in the training set, thus enhancing our model's training procedure.

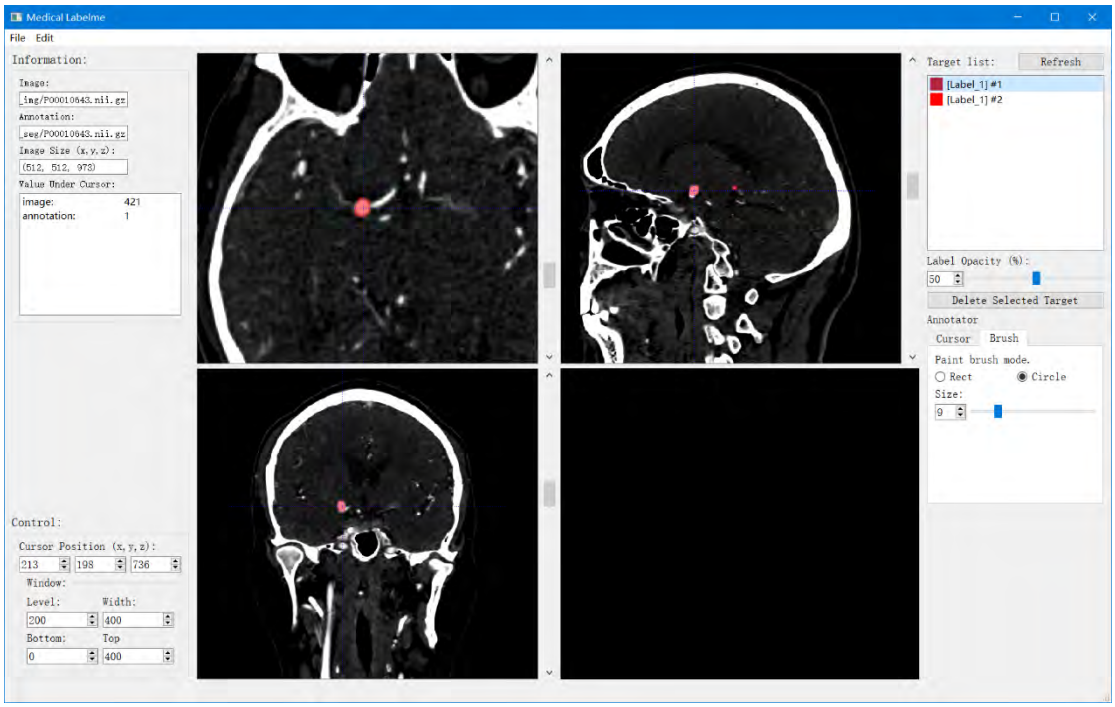


Figure S6. CTA viewing and annotation tool to assist radiologists in the IA diagnosis procedure. The software supports common scan viewing functions like the adjustment of the HU window and image statistics display. It also offers functions like IA annotation and target navigation, which is helpful to deal with large 3D images.

Supplemental Experimental Procedures

Implementation Details of GLIA-Net

Model Structure

Our segmentation model consists of a global positioning network and a local segmentation network whose inputs are the resized global CTA image and the local image patch with the same size of 96x96x96. They share some similar basic blocks in the building design. We use residual blocks[1] as the unit blocks in all architectures, which has 3 3D-convolution layers and a residual connection. All the convolution layers are followed by a group normalization layer[2] and a leaky relu activation layer[3]. We use bottleneck design for our residual block in which the kernel size for the first and last convolutions is 1 and that for the second is 3. Depending on the residual block, we design a universal encode block, which consists of a max-pooling layer if needs down-sampling and a residual block.

The global positioning network contains a global feature generator and a local feature generator. The global feature generator takes the resized global CTA image as input and has 5 encode blocks to build the global feature map whose output channels are 8, 16, 32, 64, and 128 separately, the 2nd and 3rd of which use down-sampling. Then the global feature map is cropped by a roi-pooling layer and reshaped to 6x6x6 whose bounding-box is the position of the current local patch. Finally, the local feature generator will be applied, which contains 2 encode blocks with 64 and 32 output channels. This local feature of the global positioning network for the current patch will be used to (1) compute the global positioning loss and (2) guide the local segmentation network through skip-connections. In (1), there are a 3D-convolution layer with group normalization and leaky relu, a global max-pooling layer, and a fully connected layer before the softmax computation of the global positioning loss. In (2), the output feature of the global positioning network is average-adaptive-pooled to different sizes in different scales of the local segmentation network and is activated by a sigmoid layer after a 3D-convolution layer.

The local segmentation network uses the encoder-decoder design with skip-connections between them, like U-net[4]. The encoder consists of 4 encode blocks with output channels 16, 32, 64, and 128. Except for the 1st encode block, all the other blocks contain down-sampling. So, the output feature map sizes of them are 96x96x96, 48x48x48, 24x24x24, and 12x12x12. The skip-connections is composed of the output feature maps of the first 3 encode blocks and are element-wise multiplied by specific adapted-sized local feature maps of the global positioning network. Then the enhanced skip-connections are conveyed to the decoder. The decoder of our local segmentation network consists of 3 decode blocks with output channels 64, 32, and 16 that can restore the feature map size to the original size of the input image step by step. The decode block takes the output of the former decode block (output of the encoder for the first decode block) and a skip-connection feature as input. It contains a 3D-transposed-convolution layer and a residual block. Then a final convolution layer whose output channel is 2 is applied to generate the final local segmentation probability map. The pipeline of building

pyramid weighted loss is shown in Figure S4.

Input Transformation

All the CTA images in our dataset are loaded as 3D images. The resolution of original images is Dx512x512, where D indicates the number of 2D images in each CTA scan. We clip the HU (Hounsfield unit scale) value of the images into 3 input channels before sending them to the network, each with a range of 0-100, 100-200, and 200-800. All the values in the 3 input channels are then normalized to 0-1. The clipping strategy is inspired by the diagnostic procedure of clinical practice.

The global image is resized from the original CTA image to 96x96x96 while keeping the same aspect ratio (with zero-padding) before fed into the global positioning network. The local image with a size of 96x96x96 is cropped from the original CTA image using a 3D tiling method. In clinical usage, we use a sliding window to generates local image patches from the global image with an overlap of 64 voxels, making sure that no possible target is lost. But when training, because of the severe label unbalance, we collect the training patches into a positive group and a negative group with the same numbers. For the positive group, we locate all lesion region centers and extract patches with a random deviation that is a maximum of 38 pixels from the center points, together with data augmentation of random flipping and rotation. For the negative group, we randomly select the patch centers from the global image.

Training Details

We train our end-to-end model using a deep learning framework PyTorch on RTX2080ti with 11GB memory. Adam optimizer is adopted with an initial learning rate of 0.0002 and the learning rate is decayed by 0.95 every 10000 steps. The training batch size is set to 3 and we train the model for about 200k steps. ω_{Global} and ω_{Local} in the total loss function of the training is set to 0.1 and 1.0. ω_{Dice} and ω_{Cross} in the local loss are 0.8 and 0.2. γ_{Dice} and γ_{Cross} in cross-entropy loss are both 0.3. In the cross-entropy loss of local loss, the pyramid weight for targets larger than 400 voxels is set to 3.0~20.0, and that for small targets is fixed to 11.5. The loss weight for negative voxels is set to 1.0.

Implementation Details of Other Methods

U-Net

Because the memory consumption in the 3D convolution network is heavy, the original U-net cannot be transferred to a 3D version while keeping the same parameter scale. We modify the original U-net to a similar parameter scale to our model. The encoder and decoder use the same structure as our model, except that it has no global feature to guide the skip-connections. This modified U-net uses batch-normalization and softmax cross-entropy loss to train.

HeadXNet

We follow the model structure described in the HeadXNet paper[5]. Because the specific model structure like the output channels for each block is not given, we design it to fit a similar parameter scale as ours. The HeadXNet model takes the same local images as ours and the training batch size is also 3. The output channels for each encoder block are set to 8, 16, 32, 64 and the output channel for the ASPP block is 64. There is only one max-pooling layer in the model as described in the paper, and we follow it. We also test the version that all the encoder blocks have a max-pooling layer, but the performance gets worse. We use softmax cross-entropy loss in the training period because the training using the combination of dice loss and softmax cross-entropy loss always leads to an unstable result that generates all-black outputs.

Supplemental References

1. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

2. Wu, Y., and He, K. (2018). Group normalization. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19.

3. Maas, A.L., Hannun, A.Y., and Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In Proc. icml, p. 3.

4. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention, Pt Iii *9351*, 234-241.

5. Park, A., Chute, C., Rajpurkar, P., Lou, J., Ball, R.L., Shpanskaya, K., Jabarkheel, R., Kim, L.H., McKenna, E., Tseng, J., et al. (2019). Deep Learning-Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. JAMA Network Open *2*, e195600.