

Intelligent Assisted Diagnosis and Treatment System for Vascular Infarction Based on YOLO-SAM and Unique3D

Dongcai He, Quanwang Jing, Chenyang Wang, Yehua Hu, Beining Guo, Tao Xue *

North China University of Science and Technology, Tangshan Hebei, 063000, China

* Corresponding author: Tao Xue

Abstract: Cardiovascular and cerebrovascular diseases have become the leading cause of death among chronic diseases in China. To address this, an intelligent vascular plaque detection and thrombosis risk prediction system was designed, integrating image recognition and 3D modeling into a multi-stage diagnostic framework. First, the YOLO model performs real-time plaque detection within vascular images, feeding the detected regions as input to the SAM model for precise segmentation. Subsequently, the segmented regions undergo high-fidelity 3D reconstruction using Unique3D technology to restore the spatial structure of vessels and plaques. Finally, an LSTM-based machine learning algorithm is trained using the 3D morphology and clinical labels to predict the probability of plaque-induced vascular infarction, thereby completing the closed-loop auxiliary diagnosis system.

Keywords: Vascular Plaque Detection; Medical Image Segmentation; 3D reconstruction; Deep Learning; Auxiliary Diagnosis.

1. Introduction

Cardiovascular and cerebrovascular diseases rank among the world's deadliest chronic conditions, with their onset and progression closely linked to the formation and rupture of atherosclerotic plaques. Plaque rupture readily triggers thrombus formation, leading to severe consequences such as myocardial infarction or stroke, posing a significant threat to public health. Therefore, achieving early identification and risk assessment of vascular plaques is crucial for reducing the incidence of cardiovascular and cerebrovascular events.

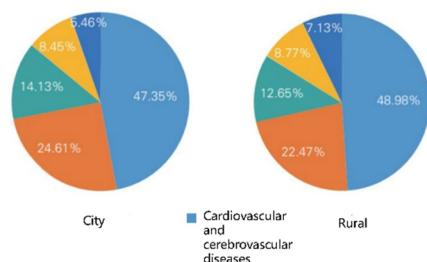


Fig 1. Proportion of Major Causes of Death Among City and Rural Residents in China

Currently, clinical diagnosis of vascular plaques primarily relies on manual interpretation of two-dimensional medical imaging such as ultrasound and CTA. This approach exhibits significant limitations, including heavy dependence on physician experience, low efficiency, and difficulties in quantitatively analyzing plaque spatial structures. Two-dimensional imaging inherently struggles to accurately represent the three-dimensional morphology of vessels and plaques, their spatial relationships, and potential risks, thereby hindering precise prediction of vascular infarction risk and personalized diagnosis and treatment.

With the advancement of deep learning technologies, artificial intelligence has demonstrated promising applications in medical image analysis. Object detection models like YOLO enable rapid lesion localization, while SAM models achieve high-precision pixel-level segmentation under guidance. However, existing research predominantly

focuses on two-dimensional detection and segmentation, lacking systematic modeling and analysis of the three-dimensional vascular-plaque structure and its clinical risks.

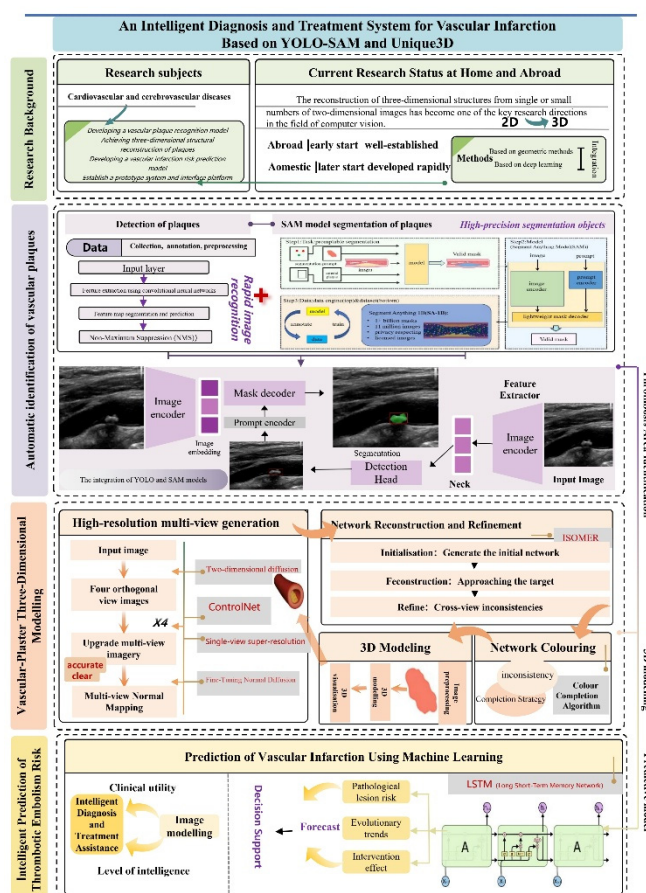


Fig 2. Technical Roadmap Flowchart

To address these challenges, this paper proposes an intelligent vascular infarction diagnosis and treatment system integrating YOLO-SAM with Unique3D. The system employs YOLO to guide SAM for efficient plaque detection

and precise segmentation, then utilizes Unique3D for three-dimensional reconstruction of vascular and plaque regions. Building upon this foundation, it combines three-dimensional structural features with clinical information to construct a risk prediction model, forming a closed-loop intelligent diagnosis and treatment process encompassing “detection-modeling-prediction.” This research contributes to enhancing the automation and precision of vascular plaque analysis, providing technical support for early screening and decision-making assistance in cardiovascular and cerebrovascular diseases.

2. Current State of Research at Home and Abroad

The SAM model was first proposed and officially released by Meta AI on April 5, 2023. MedSAM[1] is an enhanced version designed for medical image segmentation tasks. By utilizing a multimodal dataset containing over a million pairs of medical images, this model significantly improves segmentation accuracy and generalization capabilities. Compared to international research, domestic studies on the SAM model started relatively late but have developed rapidly in recent years. FastSAM[2], a lightweight image segmentation model proposed by the Image and Video Analysis Laboratory at the Chinese Academy of Sciences' Institute of Automation in 2023, employs a CNN+FPN feature extraction network within the YOLOv8-seg architecture. This replaces the computationally intensive ViT model used in SAM, enabling FastSAM to achieve comparable performance while delivering 50 times faster execution speed. Driven by large-scale data pre-training, the Multimedia Laboratory team at The Chinese University of Hong Kong proposed a zero-shot personalized SAM approach in 2023, termed PerSAM[3]. Given only a single image with a reference mask, PerSAM first locates target concepts via prior positional information, then segments them into other images or videos using three techniques: target-guided attention, target semantic prompts, and cascaded post-refinement. The Counting-SAM[4] model, proposed by Harbin Institute of Technology (Shenzhen) and Fudan University, explores SAM's application in the challenging task of small-sample object counting, which involves counting objects of unseen categories by providing bounding boxes of some examples. SAM-Adapter[5] integrates domain-specific information or visual cues into segmentation networks using a simple yet effective adapter, significantly enhancing SAM's performance on challenging tasks such as camouflaged object detection, shadow detection, and polyp segmentation (medical image segmentation). Two experiments conducted by Zhejiang University on breast tumor and maternal pelvis datasets validate that large models exhibit superior generalization capabilities after extensive fine-tuning on large datasets. Shandong University of Traditional Chinese Medicine investigated SAM's application across five major medical image modalities—whole-slide imaging, magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and multimodal imaging—by examining dataset characteristics, structural features, and generalization capabilities[6], aiming to optimize SAM's utility in medical fields.

The YOLO model is a real-time object detection algorithm. Unlike traditional methods, YOLO treats object detection as a regression problem, requiring only a single forward pass to

detect all objects in an image, significantly boosting detection speed. In 2024, YOLOv11[7] introduced an enhanced feature extraction module that more effectively captures detailed information in images, thereby improving detection accuracy. While maintaining or improving detection performance, it reduces model parameters and computational complexity, enhancing resource efficiency. This makes it one of the most advanced versions currently available. Although domestic research started relatively late, it has progressed rapidly in recent years. Many scholars have focused on adapting YOLO to specific scenarios, such as addressing the challenge of densely distributed small objects (e.g., cells) in remote sensing images[8]. They proposed improved YOLO architectures and integrated attention mechanisms[9] to enhance detection accuracy. In medical imaging, lightweight network architectures[10] and knowledge distillation techniques[11] have been introduced to reduce model parameters while preserving detection performance, making it suitable for embedded device deployment. YOLO's lightweight design, multi-sensor data fusion, adaptive learning, and intelligent annotation capabilities offer significant advantages for crop recognition and pest/disease detection[12]. Beyond agricultural applications, YOLO has also made outstanding contributions to medical diagnostics. YOLO has been applied to diverse medical tasks, including lesion detection, skin lesion classification, retinal abnormality identification, cardiac anomaly detection, brain tumor segmentation, and personal protective equipment (PPE) detection. During the COVID-19 pandemic, to alleviate the diagnostic burden on physicians for confirmed cases, YOLOv5 was employed to annotate patients' ultrasound images[13], thereby assisting clinical diagnosis.

With the rapid advancement of deep learning technology, reconstructing three-dimensional structures from single or limited two-dimensional images has emerged as a significant research direction in computer vision. This technique finds extensive applications across various scenarios, including medical imaging, autonomous driving, and digital twins. Research in this field can be broadly categorized into three approaches: geometry-based methods, deep learning-based methods, and hybrid methods that combine both. Overseas research in this domain began earlier and has developed a relatively mature research framework. With the emergence of neural rendering and implicit representations, predicting voxel meshes or implicit surfaces from single RGB or medical images has significantly enhanced reconstruction accuracy and detail recovery capabilities[14]. Payal Maken and Abhishek Gupta developed a method for 3D reconstruction from 2D X-ray images to generate 3D models, which notably reduces patient radiation exposure[15]. Domestic research began somewhat later but has advanced rapidly in recent years. For instance, hollow convolutions enable multi-scale spatial feature extraction, while spatio-temporal feature extraction modules fuse spatial and temporal information to form saliency map sequences. Convolutional Long Short-Term Memory (ConvLSTM) units are employed for generating three-dimensional dynamic content[16]. Grounded in the integration of apparel and digitalization, a novel 2D-to-3D conversion method was proposed[17]. Li Jinfang and Shi Fazhong proposed a BOX display method for converting views from two-dimensional views to three-dimensional geometric models. This method automatically positions separated and recognized views according to their projection directions, transforming a single two-dimensional

view plane into multiple projection planes. It establishes projection correspondences between views, thereby intuitively representing the projected views of an entity and aiding users in visualizing its form[18].

With the continuous advancement of artificial intelligence technology, machine learning has achieved remarkable results in medical image recognition, disease prediction, and auxiliary diagnosis. Overseas applications of machine learning in thrombosis prediction utilize ML and the Caprini model to analyze genetic data and forecast thrombosis risk. While this approach enables greater personalization between cases and captures nuances that traditional risk scores may overlook, the interpretability and clarity of outputs still require evaluation to encourage continuous improvement[19]. Internationally, automated grading systems for atherosclerotic plaques based on deep residual networks have also been developed to predict thromboembolic risk. This approach is more mature and holds leading advantages in data sourcing, model diversity, cross-modal fusion, and clinical application[20]. Domestic machine learning research in thrombus prediction began relatively late. However, in recent years, with the accumulation of electronic medical record (EMR) and medical imaging data, coupled with the widespread adoption of deep learning frameworks, related studies have advanced rapidly. Furthermore, addressing the dynamic nature of thrombus formation, some teams have attempted to combine time series modeling (e.g., LSTM) with structured data[21] to achieve continuous prediction of future thrombus formation risk in postoperative or chronic disease patients. However, current domestic research still faces challenges such as insufficient data volume, lack of multi-center validation, and difficulties in model deployment and implementation[22], with most studies remaining at the laboratory validation stage. With the continuous advancement of computer science, AI technology holds promise for assisting in the diagnosis and treatment of heterogeneous syndromes like acute coronary syndrome based on big data, as well as for intelligent recognition of imaging results related to cardiovascular diseases[23]. Its clinical application not only enhances physician efficiency and reduces error rates but also enables disease feature learning for challenging cases through this technology, intelligently generating diagnostic outcomes and treatment plans to elevate healthcare standards in remote and underdeveloped regions.

3. Automatic Identification of Vascular Plaques Based on YOLO and SAM

(1) Detection of Plaques Using the YOLO Model

YOLO is a popular object detection and image segmentation model. YOLO's unique approach treats object detection as a single regression problem, directly predicting bounding boxes and class probabilities from the entire image in a single evaluation pass—hence the name “You Only Look Once.”

A key improvement in YOLOv1 is the introduction of the C3k2 module, replacing the C2f module from previous versions. C3k2 introduces parallel convolutional layers instead of a single convolutional layer, reducing redundant computations and boosting inference speed. Input images are divided into an $S \times S$ grid, with each grid cell responsible for predicting objects whose center falls within it. Each bounding box contains the position of its center relative to the grid, the width and height of the box, and confidence information

about the object it contains. Each grid also predicts the probability of various object categories, assuming an object is present.

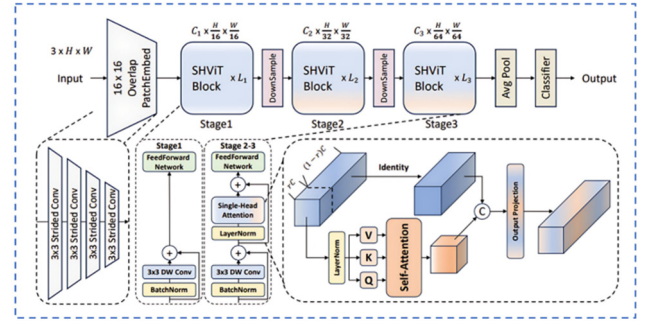


Fig 3. YOLOv11 Flowchart

Additionally, the C2PSA module enhances spatial attention, enabling the model to focus on critical regions within images. This attention mechanism potentially achieves more accurate detection, particularly for smaller or partially occluded objects, making it more supportive for plaque detection.

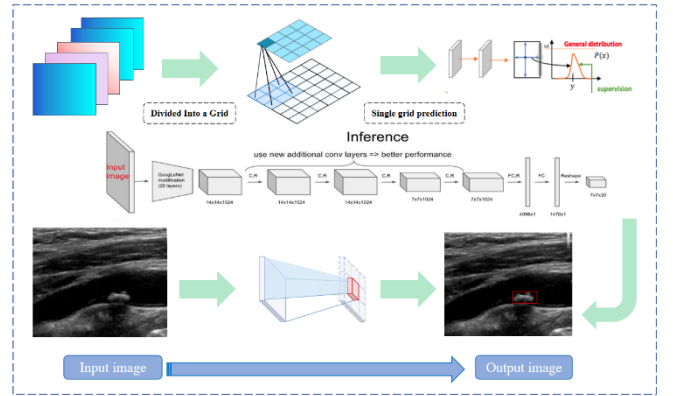


Fig 4. Test Sample Diagram

YOLOv11 employs box loss to account for factors such as box overlap, center point distance, and aspect ratio, thereby enhancing localization accuracy to bring the predicted box center as close as possible to the center of the true bounding box. The box loss formula is as follows:

$$\text{Box_Loss} = \lambda_{\text{coord}} \sum_{j=0}^S \sum_{i=0}^B 1^{obj}_{ij} ((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2) + \lambda_{\text{coord}} \sum_{j=0}^S \sum_{i=0}^B 1^{obj}_{ij} ((\sqrt{\omega_i} - \sqrt{\hat{\omega}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2) \quad (1)$$

Here, S represents the grid size, B denotes the number of predicted bounding boxes per grid cell, 1^{obj}_{ij} indicates whether the i -th bounding box in the j -th grid cell is responsible for predicting the target, and x, y denotes the width and height of the bounding box. ω, h is a weighting coefficient used to balance losses across different regions. YOLO outputs a large number of candidate boxes, which are then filtered through NMS (Non-Maximum Suppression) to eliminate duplicates, retaining only the highest-scoring boxes.

(2) Segmentation of Plaques Based on the SAM Model

The Segment Anything Model (SAM) draws inspiration from prompt strategies in the NLP field, enabling rapid segmentation of arbitrary objects by providing prompts for image segmentation tasks. SAM achieves image segmentation through multiple simple instructions. This model employs a unified architecture to process diverse instructions, directly combining prompt information with image features to generate segmentation results. Trained on large datasets containing abundant images and annotations,

and employing dynamic parameter adjustment techniques, SAM operates effectively in unseen scenarios. Experiments demonstrate its outstanding performance across multiple standard benchmarks, enabling direct application in emerging fields such as medical imaging and remote sensing. The model architecture comprises three components: an image encoder extracts image features, a prompt encoder transforms user instructions, and a mask decoder fuses both types of information to produce the final output. The technical workflow of SAM is illustrated in Figure 5.

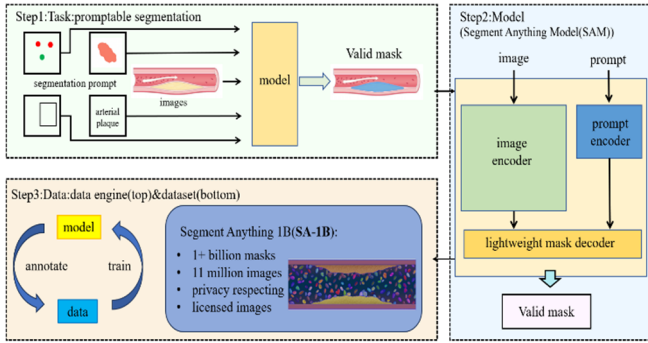


Fig 5. Technical Roadmap for the SAM Model

Image Encoder: The image encoder is essentially a stack of multiple enhanced Vision Transformer architectures. After entering the network, the original image undergoes preprocessing to resize it to 1024×1024 dimensions. Convolutional layers divide the image into small patches, converting each patch into a 768-dimensional deep feature vector. The image information is then fed into a MAE-pretrained Vision Transformer architecture for processing, ultimately outputting a feature map reorganized into a 64×64 resolution. For subsequent segmentation, the encoder incorporates a local attention mechanism and convolutional layers to highlight image features. It retains the original positional encoding while adding layer normalization and the GeLU activation function.

Prompt Encoder: Prompts can be categorized into point, bounding box, point-plus-bounding box, global, text, and other formats. The SAM model's prompt encoder converts user-inputted prompts of various types into a unified numerical representation. Point prompts are divided into foreground points and background points, marked by learnable vectors, where foreground points represent the desired regions and background points indicate areas of no interest. Boundary prompts are encoded as 8-dimensional vectors by concatenating the coordinates of the top-left and bottom-right corners after position encoding. For text prompts, the model uses a pre-trained CLIP text encoder to convert input text into 512-dimensional semantic vectors. When no user-provided prompts exist, the SAM model implicitly generates a regular grid of 1024 (32×32) points on the image as virtual prompts. It then predicts masks for the center image at each grid point, filters high-confidence masks using its built-in IoU prediction head, and finally employs 3D NMS to remove duplicate results. When mask prompts are present, the model downsamples the input low-resolution mask by a factor of 4 via convolution and aligns it with the image encoder's feature map. All types of prompt information are ultimately integrated into a 256-dimensional feature vector, combined with the image feature map through addition, and fed into the subsequent segmentation decoder.

Mask Decoder: After the image encoder in SAM generates feature maps and mask tokens, the mask decoder performs cross-attention between the 64×64 feature maps from the image encoder and the 256-dimensional feature map from the prompt encoder. This fusion is achieved through eight Transformer decoder blocks, each containing eight attention heads and a feedforward neural network. The fused features undergo 4x upsampling to generate an initial 256×256 resolution mask. Dynamic convolution modules then refine boundary details. Finally, an IoU prediction head evaluates segmentation quality, outputs two complementary masks, and generates a binary result via thresholding. Residual connections ensure information flow throughout the process.

The dataset utilizes processed medical images depicting vascular occlusion caused by venous atherosclerosis. Robustness is enhanced through data augmentation techniques, dynamic simulation, and ± 5 -pixel positional perturbations. Training employs a hybrid loss function with the AdamW optimizer across multiple epochs. This technique aims to approximate segmentation of processed medical images, though improvements remain possible for boundary and early-stage venous atherosclerosis segmentation.

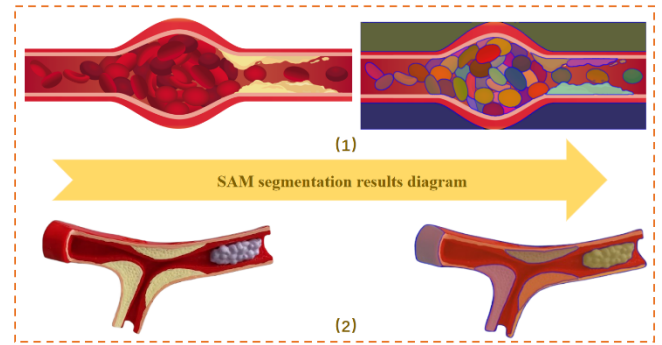


Fig 6. Sample Diagram of the Dataset

Figure 6 shows the original unprocessed image on the left and the SAM-segmented image on the right. In (1), the protruding portion on the right side of the SAM-processed image represents the segmented atherosclerotic plaque, while the central area indicates blockage caused by blood viscosity. In (2), the segmented portion in the SAM-processed image corresponds to the atherosclerotic plaque, and the right section indicates blockage caused by blood viscosity.

(3) Automatic Identification of Vascular Plaques Based on YOLO and SAM

Combining YOLO11 with SAM simultaneously enhances the capabilities of both models. YOLO excels at rapidly identifying objects within images, while SAM demonstrates superiority in high-precision object segmentation. By embedding YOLO within SAM and leveraging the complementary strengths of these two models, more comprehensive and accurate image analysis can be achieved. YOLO11 serves as the real-time detection module, incorporating a cross-scale selective fusion module. Through a bidirectional feature channel weighting fusion strategy, it significantly enhances detection capabilities for atherosclerosis and embolism. Meanwhile, SAM's medical adaptation strategy employs a multimodal prompt-driven paradigm. Prompts derived from the YOLO11 model progressively refine SAM's segmentation locations, resulting in more precise segmentation.

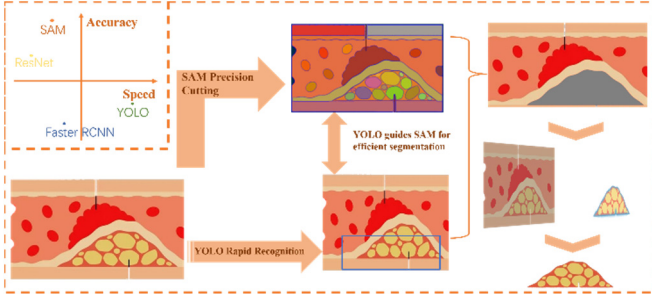


Fig 7. Schematic Diagram of YOLO Guiding SAM for Efficient Cutting

The combined approach for atherosclerosis and embolism detection proceeds as follows: YOLO11 first locates the precise positions of plaques and emboli within blood vessels. The generated detection boxes serve as input prompts for SAM, which then performs pixel-level segmentation to extract critical pathological information such as regional boundaries of plaques and emboli. The segmentation results are integrated through a decision fusion module that combines detection confidence scores with segmentation quality assessments, ultimately generating a structured diagnostic report.

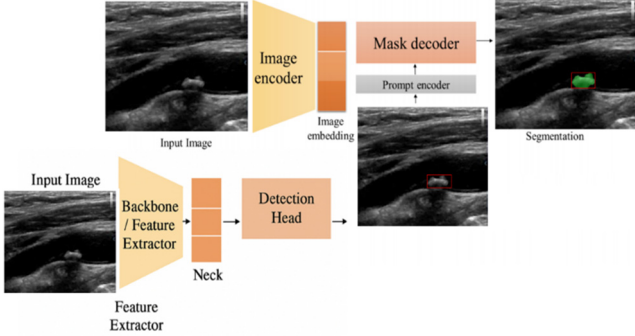


Fig 8. Example Diagram of YOLO and SAM Integration

4. Three-dimensional Reconstruction of Blood Vessels based on Unique3D

Unique3D is an image-to-3D framework that efficiently generates high-quality 3D meshes from single-view images, achieving optimal generation fidelity and strong generalization capabilities. Given an input image, it first generates four orthographic multi-view images and their corresponding normal maps using a multi-view diffusion model and a normal diffusion model. These are then progressively upscaled to high-resolution space. Finally, given high-resolution multi-view RGB images and normal maps, the timely and consistent mesh reconstruction algorithm ISOMER reconstructs high-quality 3D meshes from the high-resolution color images and normal maps. ISOMER directly addresses inconsistencies in global normals at the same vertex across different viewpoints to enhance consistency. This method achieves high-quality reconstruction and modeling for vascular lumens, plaque surface textures, and thrombus structures. The framework flowchart is shown in Figure 9, with a schematic illustration in Figure 10.

(1) High-Resolution Multi-View Generation

Design of a high-resolution multi-view generation model capable of producing four orthographic view images from a single input image. First, a pre-trained 2D diffusion model is initialized using Stable Diffusion checkpoints and fine-tuned to encode multi-view dependencies, ultimately yielding a

multi-view diffusion model that generates four orthographic view images (at 256 resolution) from a single natural image.

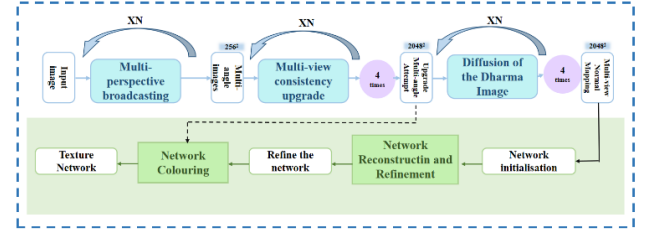


Fig 9. Example Diagram of YOLO and SAM Integration

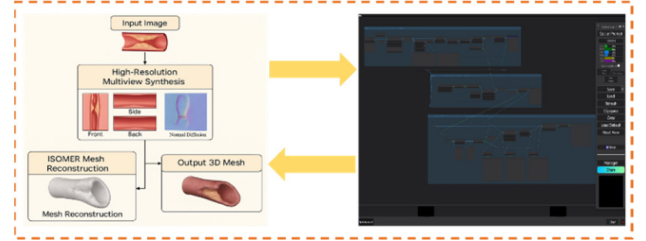


Fig 10. Unique3D Vascular Modeling Schematic Diagram

Based on the generated four orthographic view images, a multi-view-aware ControlNet is fine-tuned to enhance image resolution. This model utilizes the four juxtaposed RGB images as control information to generate correspondingly clearer and more precise multi-view results. It enhances details, improves unclear regions, and increases the image resolution from 256 to 512. Finally, a single-view super-resolution model further upscales the image by a factor of four to 2048 resolution, delivering sharper edges and finer details without compromising multi-view consistency. High-resolution normal map prediction is achieved by fine-tuning a normal diffusion model to predict normal maps corresponding to multi-view color images. This restores high-fidelity details, particularly edge precision, enhancing recovery of geometric features like plaque boundaries and vascular wall irregularities. Concurrently, a channel-level noise cancellation strategy mitigates issues arising from discrepancies between initial Gaussian noise during sampling and the most noisy training samples.

(2) ISOMER: A Highly Efficient and Consistent Grid Reconstruction Algorithm

The ISOMER algorithm's computational load scales only quadratically with spatial resolution and depends on the number of faces, enabling efficient reconstruction of meshes containing tens of millions of faces within seconds. It is particularly suited for rapid reconstruction of complex vascular systems and minute lesion structures. The process comprises three main steps: (1) Initialization: Estimates the coarse topology of 3D images and directly generates an initial network; (2) Reconstruction: Employing a coarse-to-fine strategy to progressively approximate the target shape; (3) Refinement: Explicitly resolving multi-view inconsistencies to reconstruct high-fidelity, intricate details, further enhancing the stability and realism of plaque or thrombus models. Notably, the entire mesh reconstruction process takes no more than 10 seconds. Initial mesh estimation: By directly estimating the initial mesh using front and back views, all topologically connected components visible from the front can be rapidly and accurately recovered. Specifically, we integrate the normal map of the front view to obtain:

$$d(i, j) = \sum_t = 0^i n_x(t, j) \quad (2)$$

$n_x(t, j)$ is the normal vector at the t -th pixel in the j -th row.

Although the diffusion process generates pseudo-normal maps, these are not true normal fields but rather rotation-free. To address this, the normal maps undergo random rotation prior to integration. This process is repeated multiple times, and the depth is then computed using the average of these integrations, yielding a reliable estimate. Subsequently, we map each pixel to its spatial location using the estimated depth, creating mesh models from both the front and back sides of the object. These two models are seamlessly connected via Poisson reconstruction, ensuring smooth continuity between them. Finally, they are simplified to 2000 faces for mesh initialization. Mesh optimization proceeds from coarse to fine. Based on research on inverse rendering, we iteratively refine the mesh model to minimize a loss function. At each optimization step, the mesh undergoes differentiable rendering to compute loss and gradients, followed by vertex displacement according to the gradients. Finally, the mesh is refined through iterative edge folding, splitting, and flipping to maintain uniform face distribution and reasonable edge lengths. After hundreds of coarse-to-fine iterations, the model converges to a rough approximation of the target object's shape. The loss function for this stage incorporates a mask-based loss:

$$L_{mask} = \sum_i \left\| \hat{M}_i - M_i^{pred} \right\|_2^2 \quad (3)$$

$\hat{M}_i$ is the rendered mask under view i , and M_i^{pred} is the predicted mask under view i from the previous section. The mask-based loss function adjusts the mesh contour. Additionally, it incorporates a normal-based loss function:

$$L_{normal} = \sum_i M_i^{pred} \otimes \left\| \hat{N}_i - N_i^{pred} \right\|_2^2 \quad (4)$$

For the rendered normal map $\hat{N}_i$ and predicted normal map N_i^{pred} of the object, optimize the normal direction within the visible region, $\otimes$ denotes element-wise generation. The final loss function is computed as follows:

$$L_{recon} = L_{mask} + L_{normal} \quad (5)$$

To address potential surface collapse under limited-viewpoint normal supervision, a regularization method called “expansion” is employed. At each step, vertices are displaced a small distance along their normal direction, analogous to weight decay. To mitigate severe rippling artifacts caused by all-view discrepancies, each vertex is assigned a unique optimization objective to guide directional refinement. Compared to traditional methods implicitly using multi-view images as optimization targets, the results exhibit greater robustness due to the explicit definition of the optimization objective. This explicit objective is termed *ExplicitTarget* and designed as follows: (*ExplicitTarget*). Let

$Avg(V, W) = \frac{\sum_i V_i W_i}{W_i}$ denotes the weighted average, represents the visibility v of vertex v in view i of mesh

M . $Col_M(v, i)$ denotes the color of vertex v in viewpoint i . displays objective *ET* for each vertex in mesh M is computed as follows:

$$ET_M(v) = \begin{cases} Avg(Col_M(v, i), V_M(v, i) W_M(v, i)^2) & , \text{ if } \sum_i V_M(v, i) > 0 \\ 0 & , \text{ otherwise} \end{cases} \quad (6)$$

$W_M(v, i) = -\cos(N_v^{(M)}, N_i^{(view)})$ is the weighting factor, $N_v^{(M)}$ is the vertex normal of v , and in mesh M , $N_i^{(view)}$ is the view direction of view i . Within function $ET_M(I, I_m)$, the predicted color of vertex v is computed as a weighted sum of supervised views, with weights determined by the square of the cosine angle. This is because projected area is proportional to the cosine value, and prediction accuracy is positively correlated with the cosine value. The object loss function *ExplicitTarget* for is defined as:

$$L_{ET} = \sum_i M_i^{pred} \otimes \left\| \hat{N}_i - N_i^{ET} \right\|_2^2 \quad (7)$$

N_i^{ET} is the rendered $\{ET_M(I, N^{pred}, v) | v \in M\}$ result of mesh M at the viewpoint i . The final optimized loss function is:

$$L_{refine} = L_{mask} + L_{ET} \quad (8)$$

(3) Precise coloring and efficient invisible area color completion algorithm

After generating the precise geometric structure, it must be shaded based on multi-view images. Given the inconsistencies among multi-view images, the shading process employs the same method as the refinement stage. The mesh M is colored as follows $\{ET_M(I, I_{rgb}^{pred}, v) | v \in M\}$. For filling in colors of invisible regions—such as the backside of blood vessels, occluded parts, or missing image areas—a field-based representation is introduced. These colors typically correspond to the hues of adjacent visible regions after field optimization. Specifically, traverse all vertices in the mesh to identify those lacking color from any viewpoint, marking them as “unfilled.” For each unseen vertex, locate its directly adjacent colored neighbors. If present, calculate the average of these neighbors' colors as the current vertex's color. Mark filled vertices as “colored.” Once all invisible vertices have initial colors, continue iterating several rounds of color propagation. This eliminates harsh “red-blue boundary” discontinuities, enabling smooth color transitions. This completion strategy is particularly effective for common invisible regions like the interiors of convoluted vessels and blind ends of lumens.

5. Prediction of Vascular Infarction Using LSTM-Based Machine Learning Algorithms

Building upon thrombus region identification and 3D modeling, the system incorporates an LSTM (Long Short-Term Memory Network)-based machine learning module for predicting case numbers over time, further enhancing its intelligence and clinical utility. Machine learning enables

forecasting of lesion risk, progression trends, and intervention outcomes, providing physicians with precise, data-driven decision support.

LSTM is a specialized type of recurrent neural network designed to address long-term dependency issues. Compared to standard RNNs, LSTM effectively captures long-term dependencies in time series data by incorporating “gate” mechanisms (input gate, forget gate, output gate) and cell state.

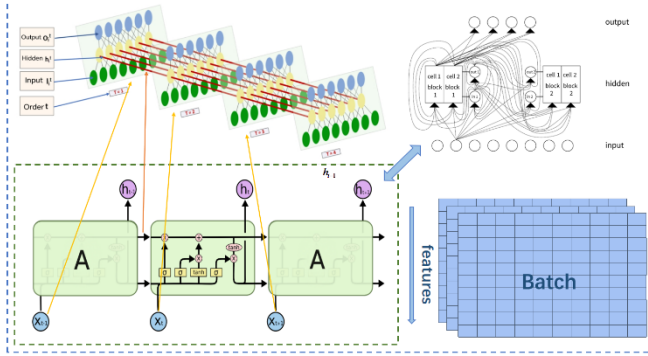


Fig 11. LSTM Schematic Diagram

LSTM incorporates a long-term memory chain and establishes connections between long-term and short-term memory chains. When computing the hidden layer state, it includes not only the input and the previous time step but also critical information from the current time step (where the input gate adds important information and the forget gate removes it). The formula is as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (9)$$

This equation represents the forget gate.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (10)$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (11)$$

This equation represents the input gate.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (12)$$

$$h_t = o_t * \tanh(C_t) \quad (13)$$

This equation represents the output gate.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (14)$$

This equation represents the cell state.

Structured image data obtained through thrombus segmentation and 3D modeling enables the extraction of core feature information for thrombus lesions at each time point. This includes geometric characteristics (such as volume, boundary area, thickness, etc.), shape characteristics (such as eccentricity, aspect ratio, curvature variation, etc.), and spatial location information (such as center-of-mass coordinates and positional changes relative to critical vascular nodes). By constructing these multidimensional features into sequential time series, the evolutionary trajectory of thrombi over time can be systematically described.

Building upon this foundation, the introduction of an LSTM (Long Short-Term Memory) based time series prediction model enables effective learning of thrombus change patterns across multiple time points. This facilitates precise prediction of thrombus status at a future specific time point or over a specific time interval. The model's output can be either classification results (e.g., whether the thrombus is

shrinking, stable, or spreading) or quantitative metrics (e.g., volume change rate, shape variation amplitude), thereby providing physicians with more detailed and reliable auxiliary diagnostic information.

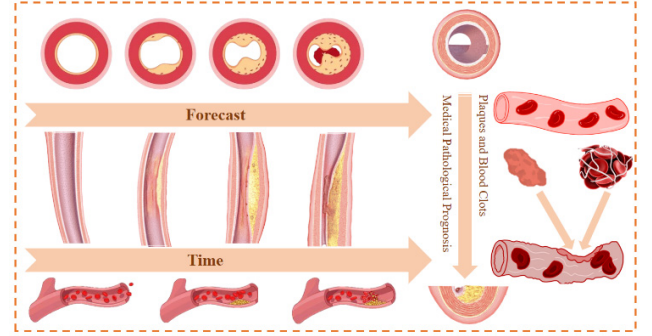


Fig 12. Prediction Diagram

This prediction mechanism not only assists in determining whether a thrombus is a structurally complex, high-rupture-risk “dangerous thrombus” and forecasts its potential to trigger acute vascular events such as cerebral infarction or myocardial infarction; it can also be used for preoperative risk assessment (judging whether thrombus is suitable for interventional treatments like thrombolysis or thrombectomy) and postoperative monitoring (predicting whether thrombus may expand, migrate, or recur). Compared to traditional static diagnostic approaches, this method significantly enhances dynamic perception of disease progression.

Overall, building upon thrombus segmentation and 3D reconstruction, this project introduces LSTM-based machine learning modeling and time-series prediction modules. This effectively achieves a leap from static image analysis to dynamic behavior modeling, representing a critical technological advancement in transitioning from “image reconstruction” to “intelligent diagnostic and therapeutic assistance.” By integrating multidisciplinary capabilities in medical image engineering, artificial intelligence, and clinical medicine, the project not only deepens and refines thrombus monitoring but also substantially enhances the system's usability, intelligence, and practical value in clinical practice.

6. Conclusion

This paper proposes an intelligent assisted diagnosis system for vascular infarction by integrating YOLO–SAM-based plaque detection and segmentation, Unique3D-based three-dimensional reconstruction, and LSTM-based risk prediction. The system forms a closed-loop framework of “detection–modeling–prediction,” effectively addressing the limitations of traditional two-dimensional image analysis in vascular disease assessment.

Results show that the YOLO–SAM collaboration improves plaque localization efficiency and segmentation accuracy, while Unique3D enables high-fidelity reconstruction of vascular and plaque structures, providing intuitive three-dimensional morphological information. Based on extracted 3D features and temporal characteristics, the LSTM model supports dynamic prediction of thrombus progression and infarction risk.

Overall, the proposed method enhances the automation, accuracy, and clinical applicability of vascular plaque analysis and risk assessment, offering a promising technical solution for intelligent diagnosis and treatment of cardiovascular and cerebrovascular diseases.

Acknowledgments

This study was supported by the Innovation and Entrepreneurship Training Program for College Students at North China University of Science and Technology, project number: S202510081018.

References

- [1] Qing Xu, Jiaxuan Li, Xiangjian He, Ziyu Liu, Zhen Chen, Wenting Duan, Chenxin Li, Maggie M. He, Fiseha B. Tesema, Wooi P. Cheah, Yi Wang, Rong Qu, Jonathan M. Garibaldi. ESP-MedSAM: Efficient Self-Prompting SAM for Universal Domain-Generalized Medical Image Segmentation. , 2024.
- [2] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, Jinqiao Wang. Fast Segment Anything. , 2023.
- [3] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, Hongsheng Li. Personalize Segment Anything Model with One Shot. , 2023.
- [4] Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan. Can SAM Count Anything? An Empirical Study on SAM Counting. , 2023.
- [5] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, Ying Zang. SAM Fails to Segment Anything? -- SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. , 2023.
- [6] Wu Tong, Hu Haoji, Feng Yang, Luo Qiong, Xu Dong, Zheng Weizeng, Jin Neng, Yang Chen, Yao Jingcao. Application of the Segmentation All Models (SAM) Approach in Medical Image Segmentation [J]. Chinese Journal of Lasers, 2024, 51(21): 1-15. DOI:10.3788/CJL240614.
- [7] Rahima Khanam, Muhammad Hussain. YOLOv11: An Overview of the Key Architectural Enhancements. , 2024.
- [8] Wang Kai, Wang Wei, Jiang Zhiwei. Detection of Small Targets in Remote Sensing Based on an Improved YOLO Model [J]. Electro-Optics and Control, 2023, 30(6): 60.
- [9] Chun-Tse Chien, Rui-Yang Ju, Kuang-Yi Chou, Enkaer Xieerke, Jen-Shiun Chiang. YOLOv8-AM: YOLOv8 Based on Effective Attention Mechanisms for Pediatric Wrist Fracture Detection., 2024.
- [10] Liu Z, Wang N, Zhang C, Wei G. [Cardiac magnetic resonance image segmentation based on lightweight network and knowledge distillation strategy]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2024 Dec 25;41(6):1204-1212. Chinese. doi: 10.7507/1001-5515.202312015. PMID: 40000210; PMCID: PMC11955363.
- [11] QinDian, BuJia-Jun, LiuZhe, ShenXin, ZhouSheng, GuJing-Jun, WangZhi-Hua, WuLei, DaiHui-Fen. Efficient Medical Image Segmentation Based on Knowledge Distillation. IEEE Transactions on Medical Imaging, 40(12): 3820–3831, 2021. DOI: 10.1109/tmi.2021.3098703.
- [12] Cao Liang, Xiao Wei, Li Xiangli. YOLO Algorithm and Its Applications in Crop Recognition and Pest Detection: A Review [J/OL]. Journal of Zhongkai Institute of Agricultural Engineering, 1-14[2025-04-20]. <http://kns.cnki.net/kcms/detail/44.1660.S.20250418.1342.002.html>.
- [13] Su Qinghua, Zhou Yufei, Qin Zhenbo, Mu Siyu, Yang Xuedong. Detection of COVID-19 Disease Based on YOLOv5 [J]. Biomedical Science, 2022, 12(4): 271-278. DOI:10.12677/hjbm.2022.124033.
- [14] XieHaozhe, YaoHongxun, SunXiaoshuai, ZhouShangchen, ZhangShengping. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. //Proceedings of the IEEE/CVF international conference on computer vision, 2019.
- [15] Maken P, Gupta A. 2D-to-3D: A Review for Computational 3D Image Reconstruction from X-ray Images[J]. Archives of Computational Methods in Engineering, 2023, 30: 85–114. DOI: 10.1007/s11831-022-09790-z.
- [16] Weng Gaoyi, Yao Jianmin, Chen Enguo, et al. Research on 2D-to-3D Conversion Technology Based on Saliency Maps [J]. Information Technology and Informatization, 2025, (02): 183-186.
- [17] Tian Yuling. Research on Visualization Conversion Technology from 2D Parameters to 3D Rendering of Turned Collars [D]. Donghua University, 2009.
- [18] Li Jinfang, Shi Fazhong. BOX Display Method in Conversion from 2D Views to 3D Geometric Models [J]. Computer Engineering, 2005, (19): 182-183+188.
- [19] I-SherbiniAdham H, CoroneosStefania, ZidanAli, OthmanMaha. Machine learning as a diagnostic and prognostic tool for predicting thrombosis in cancer patients: a systematic review. //Seminars in Thrombosis and Hemostasis, 2024.
- [20] X, Li Y, Yan F, et al. Automatic coronary plaque detection, classification, and stenosis grading using deep learning and radiomics on computed tomography angiography images: a multi-center multi-vendor study[J]. European radiology, 2022, 32(8): 5276-5286.
- [21] Yao Chunxiao. Research on a Cerebrovascular Disease Prediction System Based on Long Short-Term Memory Neural Networks [D]. Beijing Jiaotong University, 2019. DOI:10.26944/d.cnki.gbjfu.2019.000978.
- [22] Wu Li, Luo Yefangxin, Shi Wenting, et al. Establishment and Evaluation of a Machine Learning-Based Early Warning Model for Lower Limb Deep Vein Thrombosis Risk in Hemorrhagic Stroke Patients [J]. Journal of Zunyi Medical University, 2024, 47(09): 903-909. DOI:10.14169/j.cnki.zunyixuebao.2024.0118.
- [23] Lu Yanqiao, Shen Lan, He Ben. Application of Artificial Intelligence in the Auxiliary Diagnosis and Treatment of Cardiovascular Diseases [J]. Journal of Shanghai Jiao Tong University (Medical Edition), 2020, 40(2): 123-128. DOI: 10.3969/j.issn.1674-8115.2020.02.019.