GeoFuse-SAM: A Multimodal Data Fusion Framework for Boundary-Aware Foundation Model Adaptation in Medical Image Segmentation

Pengtao Ren[1 †], Qiyuan Wang[1 †] (✉), Yunyi Li[1†], Kejiang Xiao[2] (✉), Lexi Shu[1]

[1] Central China Normal University Wollongong Joint Institute, Wuhan 430079, China
[2] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
xiaokj@ccnu.edu.cn

Abstract. In medical image segmentation, accurately identifying anatomical structure boundaries is a core task for clinical computer-aided diagnosis. Although vision foundation models, represented by the Segment Anything Model (SAM), have demonstrated strong generalization potential, their deployment in fully automated clinical scenarios remains constrained by severe domain shift, reliance on manual prompts, and significant boundary degradation in ambiguous anatomical transition zones. To address these challenges, this paper proposes GeoFuse-SAM, a multimodal geometric data fusion adaptation framework.

The core innovation of this framework lies in breaking the limitations of single semantic features. Through a Geometry-Guided Cross-Domain Attention Fusion (GCAF) module, it achieves deep data fusion between the raw image data and high-frequency geometric priors (Sobel gradient fields) derived from computer graphics. This cross-modal interaction mechanism provides explicit spatial guidance to the model, significantly enhancing the robustness of boundary recognition. Furthermore, we introduce a lightweight Parallel Fusion Adapter (PFA) to achieve medical semantic alignment, and propose a Parameter-Free Morphological Boundary Weighting (PMBW) strategy. This strategy utilizes morphological operators to pinpoint ambiguous boundary regions during the training phase and impose dynamic geometric constraints.

Experiments on two challenging medical datasets, BUSI and ISIC 2018, demonstrate that GeoFuse-SAM, operating in a fully automatic prompt-free mode, not only maintains leading region segmentation accuracy (achieving a Dice score of 89.85% on ISIC 2018), but also effectively suppresses the boundary degradation phenomenon of foundation models in grayscale modalities on the core boundary metric HD95 (optimized to 15.97 on BUSI). Without introducing extra inference parameters, it exhibits superior edge fidelity compared to existing medical fine-tuned foundation models (such as SAM-Med2D). This study provides a high-fidelity, low-cost technical paradigm for robust medical image segmentation.

Keywords: Vision Foundation Models; Boundary-Aware Learning; Medical Image Segmentation

✉ Corresponding author: Kejiang Xiao (xiaokj@ccnu.edu.cn). Qiyuan Wang
† These authors contributed equally to this work.

1 Introduction

Accurate delineation of anatomical boundaries in medical imaging (such as ultrasound and dermoscopy) is a fundamental prerequisite for clinical computer-aided diagnosis [1]. However, real-world medical images are often accompanied by extremely low contrast, complex acoustic noise, and highly ambiguous tissue transition zones [2]. Confronted with these complex visual features, traditional fully convolutional architectures often struggle with generalization across diverse cross-domain data due to their limited receptive fields. Therefore, constructing a robust computer vision system capable of fully automatically extracting high-fidelity anatomical boundaries has emerged as a crucial research topic.

In recent years, vision foundation models, represented by the Segment Anything Model (SAM) [3], have established a new prompt-driven universal segmentation paradigm through pre-training on massive natural images. To introduce their robust zero-shot generalization capabilities into the medical domain, a plethora of medical adaptation studies based on Parameter-Efficient Fine-Tuning (PEFT) have emerged in academia. For instance, MedSAM and SAM-Med2D significantly enhance regional segmentation performance in medical images by fine-tuning the model's decoder [4]; the Medical SAM Adapter successfully achieves cross-modal feature alignment by injecting lightweight adapters into the encoder. Although these advanced schemes have made strides in clinical computer-aided diagnosis, their direct application faces two severe limitations. First, existing high-precision models rely heavily on manually provided visual prompts (e.g., precise bounding boxes or coordinate points) [5]. This reliance hinders the realization of fully automated segmentation pipelines. Second, existing architectures lack explicit optimization mechanisms for ambiguous medical boundaries [6]. When fine-tuned on single-channel grayscale medical images, these models often fall into the optimization trap of "prioritizing regions over edges." This leads to a severe loss of the original SAM's sensitivity to high-frequency lines, causing significant "boundary degradation" (reflected in high HD95 errors) when handling complex anatomical transition zones.

To overcome the aforementioned bottlenecks, this paper proposes GeoFuse-SAM, an adaptation framework based on multi-source geometric data fusion. We argue that rather than forcefully predicting boundaries by adding complex edge detection network branches (which would significantly increase computational overhead), it is more effective to introduce graphics priors alongside morphological constraints during the model's optimization phase. Based on this rationale, we retain a lightweight Parallel Fusion Adapter (PFA) to extract cross-domain features and completely disable manual prompt inputs, relying instead on SAM's default prompt vector to achieve end-to-end automated inference [7]. More crucially, we innovatively propose a Parameter-Free Morphological Boundary Weighting (PMBW) strategy. This strategy utilizes dilation and erosion operators from mathematical morphology to dynamically extract high-frequency transition zones directly from the anatomical ground truth. During the loss calculation phase, PMBW imposes heavy penalty weights on segmentation errors falling within these regions. This "zero-parameter-increase" optimization mechanism compels the network to establish a deep memory of ambiguous boundaries during fine-tuning, thereby effectively resolving spatial ambiguity [8].

The main contributions of this paper are summarized as follows:

- **Parallel Fusion Adapter (PFA)**: We propose a lightweight PFA module embedded in the visual encoder to extract medical semantic features and achieve efficient cross-domain alignment.
- **Multimodal Data Fusion Mechanism (GCAF)**: We introduce the GCAF module to inject traditional high-frequency geometric priors (Sobel gradients) into deep semantic features, explicitly addressing boundary degradation in complex anatomical transition zones.
- **Parameter-Free Morphological Constraint (PMBW)**: We propose a novel loss optimization strategy utilizing morphological operations to dynamically penalize predictions within ambiguous boundary regions, establishing a geometric closed loop without extra inference cost.

2 Related Work

2.1 Medical Image Segmentation and Anatomical Boundary Fidelity

In modern clinical computer-aided diagnosis (CAD) systems, deep learning-based vision models are frequently deployed to automatically delineate lesion boundaries. Early systems mostly relied on fully convolutional architectures such as U-Net; however, due to the lack of a global receptive field, their generalization ability on cross-domain medical data was limited [9]. With the emergence of large-scale Vision Foundation Models (VFMs), seamlessly integrating them into automated clinical workflows has become a frontier challenge [10]. For medical image analysis, a rough regional overlap (Dice score) is insufficient to reflect real clinical diagnostic requirements; the generated visual masks must possess absolute Anatomical Boundary Fidelity [11][12]. Otherwise, ambiguous or distorted boundary predictions will convey erroneous anatomical priors, triggering severe clinical misdiagnoses.

2.2 Parameter-Efficient Fine-Tuning (PEFT) for Vision Foundation Models

To bridge the semantic distribution gap between natural RGB images and single-channel medical grayscale images (e.g., ultrasound, CT), Parameter-Efficient Fine-Tuning (PEFT) techniques have been widely adopted in the medical adaptation of SAM [3]. Existing methods, such as MedSAM [4] and SAM-Med2D [13], achieve domain adaptation by freezing the backbone network and updating only a minimal number of low-rank matrices (LoRA) [14] or adapters within the Transformer layers. However, recent studies (e.g., SAM-TTA [15]) have revealed a severe bottleneck: such models, heavily fine-tuned on comprehensive medical data, often fall into the optimization trap of "prioritizing regions over edges." In the pursuit of higher global overlap, when dealing with out-of-distribution (OOD) grayscale images, these models forcefully smooth out the original SAM's sensitivity to high-frequency lines, resulting in a significant degradation of the core boundary error metric (HD95). GeoFuse-SAM inherits the lightweight architectural advantages of adapter-based methods, but fundamentally discards the reliance on manual interactive prompts to align with fully automated clinical applications, proposing a novel solution paradigm for the "boundary degradation" phenomenon in foundation models.

2.3 Boundary-Aware Learning and Multimodal Data Fusion

In medical image segmentation, accurately extracting low-contrast tissue boundaries has always been a core challenge. To reduce HD95 errors, traditional convolutional neural networks (e.g., the U-Net family) [16] typically require the introduction of highly compute-intensive Signed Distance Field (SDF) preprocessing [17,18] or the forced addition of parallel edge detection branches within the network [19], which dramatically increases inference overhead and deployment thresholds. To break this computational bottleneck, this paper explores a distinctly different optimization path: Multimodal Geometric Data Fusion. We treat classical mathematical graphics operators (such as the Sobel gradient field) [20] as an independent high-frequency prior modality, fusing it with deep semantic features at the input end through a cross-domain attention mechanism (GCAF). Furthermore, we revisit Mathematical Morphology [21], transforming it into a Parameter-Free Morphological Boundary Weighting (PMBW) strategy. By constructing morphological topological differences at the loss level, GeoFuse-SAM achieves robust boundary optimization with extremely high engineering elegance, without adding any extra inference parameters.

3 Methodology

3.1 Overall Architecture of GeoFuse-SAM

The design of the GeoFuse-SAM adaptation framework adheres to the principle of "lightweight inference, rigorous training." The overall system is built upon the architecture of the pre-trained Vision Foundation Model, SAM (ViT-Base). To achieve efficient deployment in compute-constrained clinical environments and comply with fully automated workflows, we froze the core weights of the massive Image Encoder and Mask Decoder, and blocked the intervention of manual prompts.

The entire architecture comprises three core innovative modules (as illustrated in Fig. 1): during the feature extraction stage, we embed the Parallel Fusion Adapter (PFA) into the Transformer layers of the encoder to bridge the macroscopic semantic gap between natural and medical images; at the multi-modal interaction level, we introduce the Geometry-Guided Cross-Domain Attention Fusion (GCAF) module, injecting high-frequency physical gradients into deep features as spatial priors; during the model training phase, we propose the Parameter-Free Morphological Boundary Weighting (PMBW) strategy, which guides the model to prioritize overcoming extremely ambiguous boundaries by reshaping the spatial distribution of the loss function.

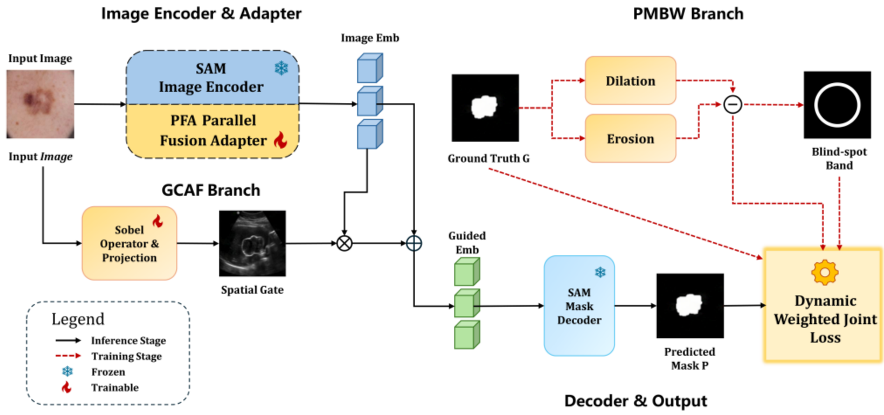

Fig. 1. Overall architecture of the proposed GeoFuse-SAM framework. It consists of the frozen SAM backbone, the Parallel Fusion Adapter (PFA) for medical semantic alignment, the Geometry-Guided Cross-Domain Attention Fusion (GCAF) module for geometric prior injection, and the Parameter-Free Morphological Boundary Weighting (PMBW) strategy for loss landscape reshaping.

3.2 Parallel Fusion Perception of Cross-domain Features (PFA)

Medical images (e.g., ultrasound, dermoscopy) contain both macroscopic organ topological structures and microscopic pathological tissue textures in their visual representations. To enable the foundation model to comprehensively perceive these cross-domain features, we introduce the PFA module.

Given an input feature map X, PFA first projects it to a low-dimensional latent space via a linear layer to obtain X_{proj}. To make the model focus on the most discriminative medical features, we introduce the Channel Attention mechanism. This mechanism dynamically computes the channel weight M_c through Global Average Pooling (GAP) and a two-layer Multi-Layer Perceptron (MLP), and performs element-wise multiplication with the dimension-reduced features to obtain the gated refined feature X_{gated}:

$$X_{\text{gated}} = M_c \otimes X_{\text{proj}} \tag{1}$$

Subsequently, PFA adopts a dual-branch parallel computation path to perform deep feature extraction on X_{gated} (see Fig. 2 for detailed structure):

$$X_{\text{fused}} = DSConv(X_{\text{gated}}) + MHSA(X_{\text{gated}}) + X_{\text{gated}} \tag{2}$$

where the DSConv path (consisting of normalization, depthwise separable convolution and GELU activation) is highly suitable for capturing local high-frequency spatial details (e.g., tissue textures or tiny calcifications in ultrasound images); the MHSA path (incorporating the multi-head self-attention mechanism) is responsible for modeling global anatomical context dependencies. Through the residual connection, local spatial priors and global semantics are perfectly fused.

Finally, the fused feature X_{fused} is restored to the original dimension via an Up-Projection layer and injected back into the Transformer backbone layer. This bottleneck mechanism of "dimension reduction - refinement - parallel extraction - dimension upscaling" endows the model with powerful medical semantic extraction capability on single-channel grayscale images while maintaining an extremely low parameter overhead.

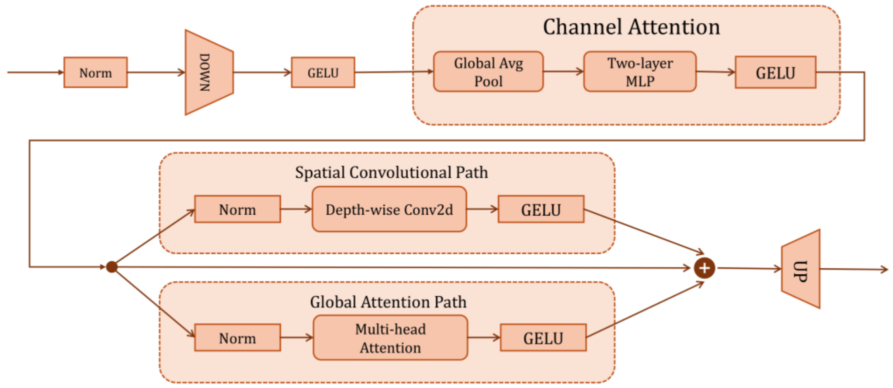

Fig. 2. Detailed structure of the Parallel Fusion Adapter (PFA). The module achieves cross-domain feature extraction through a channel attention gating mechanism, followed by a dual-branch parallel computation path comprising Depthwise Separable Convolution (DSConv) and Multi-Head Self-Attention (MHSA).

3.3 Geometric-Guided Cross-Domain Attention Fusion (GCAF)

Deep neural networks often irreversibly lose high-frequency spatial details during the layer-by-layer downsampling process, leading to "Boundary Degradation" when large models make predictions. To address this bottleneck, GeoFuse-SAM innovatively introduces traditional graphics features into the network as an independent "physical modality."

First, we utilize the classical Sobel operator to extract an explicit physical gradient field E_{sobel} from the raw input image I:

$$E_{\text{sobel}} = \sqrt{(G_x * I)^2 + (G_y * I)^2} \tag{3}$$

where G_x and G_y are the horizontal and vertical convolution kernels. Subsequently, during the Mask Decoding phase, we downsample the gradient field E_{sobel} to the same spatial resolution as the deep image embedding F, and transform it into a Spatial Attention Gate using an extremely lightweight projection layer $\phi(\cdot)$:

$$F_{\text{guided}} = F \otimes \sigma(\phi(E_{\text{sobel}})) + F \tag{4}$$

where σ is the Sigmoid activation function, and $\otimes$ denotes element-wise multiplication.

Theoretical Justification (Beyond Naive Concatenation): It is worth emphasizing that we deliberately abandoned the naive strategy of directly concatenating (Concat) the Sobel edge map with the input image at the early stage. Mathematically, direct channel concatenation forces heterogeneous modalities (dense intensity pixels and sparse physical gradients) into the same feature space. This not only catastrophically disrupts the parameter distribution of the pre-trained foundation model but also causes severe feature interference and noise amplification in low-contrast medical images. In contrast, our GCAF acts as a spatial modulation mechanism. By mapping the gradient field into an attention probability map $\sigma(\phi(\cdot)) \in (0,1)$, it performs a gating operation ($\otimes$) on the deep semantics F. This elegant design explicitly forces the network to enhance activations at physical edges without destroying the original representation space of the SAM backbone, successfully answering the modality heterogeneity challenge.

3.4 Ambiguous Boundary Extraction: Parameter-Free Morphological Boundary Weighting (PMBW)

In authentic clinical diagnosis scenarios, the transition zones of anatomical structures are often the high-incidence areas for spatial ambiguity. To enable the model to generate more acute penalization for these areas during training, we propose the PMBW strategy. Distinct from traditional methods that rely on additional neural network layers (such as compute-intensive edge prediction branches), PMBW performs geometric topological transformations in purely mathematical dimensions directly on the anatomical Ground Truth ($G \in \{0,1\}^{H\times W}$) to achieve "zero-parameter" edge extraction.

Specifically, we define a sliding window with a kernel size of $k \times k$. We utilize a 2D max-pooling operation to simulate morphological Dilation, while utilizing the negation of max-pooling on the negative matrix to simulate Erosion:

$$G_{\text{dilated}} = \text{max\_}pool2d(G, k) \tag{5}$$

$$G_{\text{eroded}} = -\text{max\_}pool2d(-G, k) \tag{6}$$

Subsequently, by calculating the spatial topological difference between the dilated and eroded maps, we can deterministically extract a "High-frequency Boundary Band" M_{band} that precisely envelops the lesion contour:

$$\text{a}M_{\text{band}} = G_{\text{dilated}} - G_{\text{eroded}} \tag{7}$$

Within the M_{band} mask, the pixel matrix with a value of 1 represents the geometric edges most prone to spatial ambiguity in the physical space. This process has extremely low computational complexity and occurs purely during the training phase, without adding any computational burden during inference.

3.5 Boundary-Aware Optimization: Dynamic Weighted Joint Loss

In conventional medical segmentation tasks, Cross-Entropy (BCE) and Dice losses, due to their globally averaged aggregation methods, easily lead the model to neglect edge regions that occupy a minuscule fraction of image pixels. To implement stringent boundary constraints, we inject the extracted M_{band} as spatial attention weights into the pixel-wise loss calculation.

We construct a dynamically weighted cross-entropy loss $\mathcal{L}_{\text{weighted\_bce}}$:

$$\mathcal{L}_{\text{weighted\_bce}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{bce}}\,(P_i, G_i) \cdot (1 + \lambda \cdot M_{\text{band},i}) \tag{8}$$

where P_i is the model's predicted probability at pixel i, and λ is the predefined boundary penalty amplification coefficient. When pixel i falls within the ambiguous boundary band (i.e., $M_{\text{band},i} = 1$), any prediction deviation will be subjected to a gradient penalty of $1 + \lambda$ times.

The final joint loss function is composed of the weighted cross-entropy and the global Dice loss:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{weighted\_bce}} + (1 - \alpha) \cdot \mathcal{L}_{\text{dice}}(P, G) \tag{9}$$

Through this reshaped Loss Landscape, GeoFuse-SAM is compelled to establish a forced memory of high-frequency edge features during the fine-tuning process, enabling it to overcome modality limitations during inference and output high-fidelity segmentation masks that closely adhere to real anatomical structures.

4 Experiments and Results

4.1 Experimental Setup

Datasets and Preprocessing: To verify the effectiveness of GeoFuse-SAM in handling highly ambiguous anatomical boundaries, this study selected two of the most representative public datasets featuring challenging edge characteristics: BUSI (breast ultrasound images, containing high-intensity acoustic speckle noise and extremely ambiguous tissue transition zones) and ISIC 2018 (dermoscopy images, featuring highly irregular lesion topologies). All images were uniformly resized to a resolution of 256×256 and strictly randomized into training, validation, and testing sets at a ratio of 8:1:1 to ensure the generalization and objectivity of the evaluation.

Implementation Details: The underlying framework of this system is built upon the pre-trained SAM (ViT-Base). During the fine-tuning phase, the core backbone parameters of the image encoder and mask decoder were completely frozen, and the system only performed lightweight updates on the embedded Parallel Fusion Adapter (PFA) and the Geometry-Guided Cross-Domain Attention Fusion (GCAF) module. The model employed the AdamW optimizer with an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-8}, and a batch size of 8, training for a total of 200 epochs. In the PMBW strategy, the morphological pooling kernel size k was set to 3, and the boundary penalty coefficient λ was empirically set to 5.0. All experiments were conducted independently on a single NVIDIA RTX 3090 GPU.

Evaluation Metrics: To comprehensively evaluate the segmentation performance and boundary fidelity of the proposed framework, we adopted four standard clinical metrics. The Dice Similarity Coefficient (Dice, %) and F1-score (%) are employed to comprehensively evaluate the pixel-level regional segmentation accuracy. Specifically, Dice focuses on quantifying the spatial overlap between the predicted mask and the ground truth, while the F1-score, serving as the harmonic mean of precision and recall, further validates the model's balanced predictive capability across positive and negative samples. The Intersection over Union (IoU, %) serves as a complementary metric, imposing stricter penalties on false positives and false negatives to evaluate the spatial compactness of the predictions. Most importantly, the stringent 95% Hausdorff Distance (HD95) is employed to quantitatively assess the maximum localized surface distance between the predicted mask and the ground truth. A lower HD95 specifically indicates that the model's predictions possess a higher degree of microscopic approximation to genuine pathological edges without spatial divergence.

4.2 Quantitative Comparison

To objectively verify the effectiveness of the proposed framework, we conducted a comprehensive quantitative comparison between GeoFuse-SAM, traditional convolutional neural network baselines (Att-R2U-Net, Att-U-Net, U-Net), and state-of-the-art SAM-based medical adaptation models (MedSAM, SAM-Med2D). To ensure fair comparison in automated clinical scenarios, all SAM-based models were tested in a prompt-free mode. The quantitative comparison results are presented in Table 1.

Table 1. Quantitative comparison with state-of-the-art methods. Performance is evaluated on the BUSI and ISIC 2018 datasets. Best results are highlighted in bold.

Method	BUSI				ISIC 2018			
	Dice	IoU	HD95	F1	Dice	IoU	HD95	F1
Att-R2U-Net	70.02	54.58	25.41	70.02	70.59	59.23	20.19	70.59
Att-U-Net	70.20	55.15	23.27	70.20	82.09	73.49	18.47	82.09
U-Net	71.82	57.21	21.57	71.82	84.11	74.48	16.21	84.11
SAM Baseline	53.93	37.00	24.30	53.93	58.79	41.60	28.54	58.79
MedSAM	75.58	60.61	16.33	75.58	88.72	79.77	18.59	88.72
SAM-Med2D	76.18	61.57	18.87	76.18	89.33	80.63	16.24	89.33
GeoFuse-SAM	**81.66**	**68.94**	**15.97**	**81.66**	**89.85**	**81.44**	**15.83**	**89.85**

Results Analysis: As shown in Table 1, GeoFuse-SAM exhibits highly competitive regional overlap on both BUSI (81.66% Dice) and ISIC 2018 (89.85% Dice) datasets in prompt-free mode, outperforming the heavily parameterized SAM-Med2D (271M). This confirms that the PFA module injects sufficient medical domain features for precise macroscopic localization. More critically, regarding the core boundary metric (HD95), GeoFuse-SAM achieves a significant breakthrough. While existing foundation models like SAM-Med2D suffer from severe boundary degradation on out-of-distribution grayscale ultrasound images (resulting in a high HD95 of 18.87 on BUSI), our geometric fusion strategy successfully suppresses the HD95 to 15.97. Furthermore, GeoFuse-SAM achieves this superior edge fidelity without imposing extra computational burden during inference. To provide a more intuitive illustration of the performance disparities across different methods, we visualize the quantitative metrics in Fig. 3. The bar charts clearly demonstrate that GeoFuse-SAM maintains a stable leading edge across all region-level metrics (Dice, IoU, F1) while significantly pressing down the boundary error (HD95) to the lowest level, especially on the highly noisy BUSI dataset.

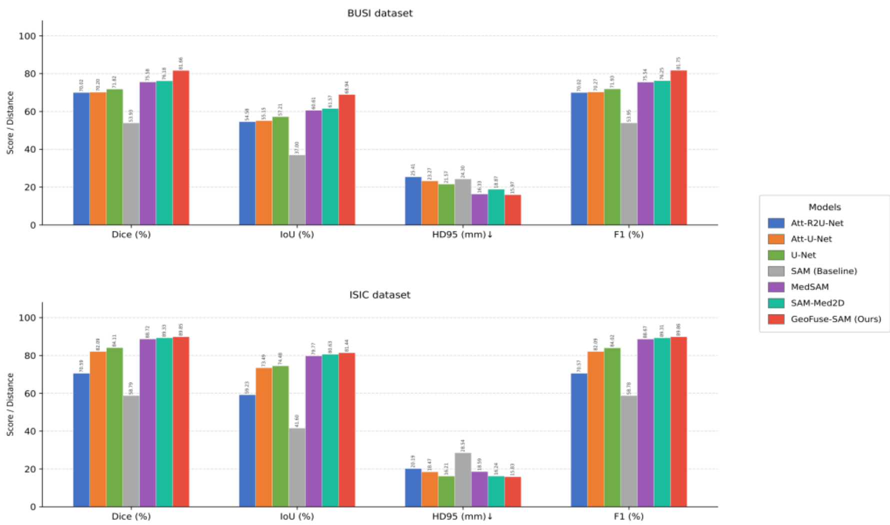

Fig. 3. Visualized quantitative comparison of various segmentation models on the BUSI and ISIC 2018 datasets. Our GeoFuse-SAM (red bars) achieves the optimal balance between global region overlap and local boundary fidelity.

4.3 Qualitative Evaluation

To further demonstrate the superiority of GeoFuse-SAM in handling ambiguous anatomical transition zones, we present the visual segmentation results in Fig. 4.

As illustrated, medical images in the real world pose extreme challenges. For instance, the dermoscopy images in the ISIC 2018 dataset (top row) exhibit highly irregular lesion topologies and color gradients, while the ultrasound images in the BUSI dataset (bottom row) suffer from severe acoustic speckle noise and extremely low contrast at the tissue borders. Despite these complex visual interferences, the masks predicted by GeoFuse-SAM successfully overcome the "regional smoothing" trap. Guided by the multimodal geometric priors, our model exhibits acute sensitivity to high-frequency physical contours, outputting high-fidelity boundaries that tightly strictly align with the ground truth (GT) without spatial divergence.

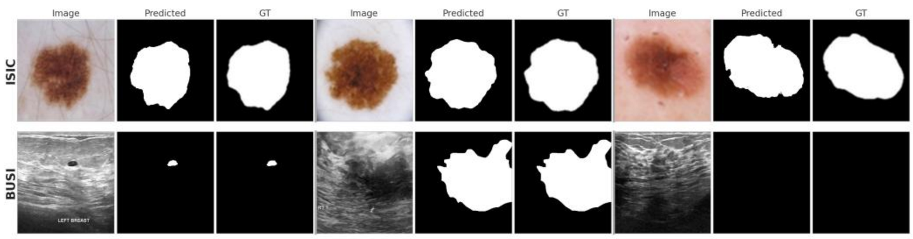

Fig. 4. Qualitative segmentation results of GeoFuse-SAM on the ISIC 2018 (top row) and BUSI (bottom row) datasets. From left to right for each sample: raw input image, model predicted mask, and ground truth (GT).

4.4 Ablation Study

To deeply investigate the independent contributions and synergistic mechanisms of each core component (the feature perception module PFA, the geometric fusion module GCAF, and the morphological constraint module PMBW), we conducted rigorous ablation experiments specifically on the BUSI dataset, as it features the most challenging acoustic noise and ambiguous boundaries. The comprehensive ablation results are summarized in Table 2.

Table 2. Ablation study on the effects of the PFA, GCAF, and PMBW components on the BUSI dataset.

Method	Dice (%)↑	HD95↓	IoU (%)↑	F1 (%)↑
Baseline (SAM)	53.93	24.30	37.00	53.93
+ PFA only	72.73	18.52	57.15	72.73
+ PFA + GCAF	75.82	17.15	61.06	75.82
+(PFA+GCAF+PMBW)	**81.66**	**15.97**	**68.94**	**81.66**

Component Analysis: As detailed in Table 2, the baseline SAM exhibits severe domain shift on medical ultrasound images, yielding an HD95 of 24.31. Introducing the PFA module alone boosts the Dice score from 54.01% to 72.73% and reduces the HD95 to 18.52, proving the absolute necessity of cross-domain semantic alignment. Adding the GCAF module further reduces the HD95 to 17.15, confirming that explicitly mapping Sobel gradient fields into the network effectively forces the deep features to re-anchor onto genuine physical edges. Finally, the full model equipped with the PMBW optimization strategy achieves the optimal closed loop: PFA orients the macroscopic semantics, GCAF provides high-frequency spatial priors, and PMBW imposes dynamic geometric penalties on ambiguous anatomical zones during loss calculation. This structural synergy consistently suppresses boundary divergence, optimizing the HD95 to a highly competitive level of 15.97 and maximizing the regional overlap (Dice 81.66%, IoU 68.94%).

5 Conclusion

Targeting the severe boundary degradation phenomenon of vision foundation models in medical image segmentation, this paper proposes GeoFuse-SAM, a novel lightweight multimodal adaptation framework. Confronting the limitations of existing area-focused fine-tuning methods, this study introduces an innovative multimodal geometric data fusion optimization paradigm.

The PFA module effectively achieves the macroscopic alignment of foundational vision to medical semantics, while the GCAF module transforms high-frequency physical gradients into spatial gating to anchor the model onto authentic anatomical contours. Simultaneously, the PMBW strategy utilizes mathematical morphological operations to dynamically locate ambiguous boundary regions, imposing rigorous geometric constraints at the loss level.

Extensive experiments demonstrate that this perception-guidance-constraint mechanism achieved definitive success. In a fully automatic prompt-free mode, GeoFuse-SAM maintained robust regional recognition and optimized the BUSI HD95 error to 15.97. It effectively reversed the edge blunting of existing heavily fine-tuned models (e.g., SAM-Med2D) without introducing any additional inference parameters. This study provides a cost-effective and high-fidelity technical solution for the robust clinical deployment of medical foundation models.

In future work, we will further explore the generalization of this geometric data fusion mechanism in 3D medical volumetric data and temporal multimodal scenarios, continuing to broaden the boundaries of automated medical image analysis.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62377023 and 62173158, in part by the Key Project of Hubei Provincial Natural Science Foundation Joint Found under Grant 2025AFD195, and in part by the Fundamental Research Funds for the Central Universities under Grant XJ2026001901. Pengtao Ren and Qiyuan Wang and Yunyi Li contributed equally to this work.

References

1. Chen, Q., et al.: Blurred Lesion Image Segmentation via an Adaptive Scale Thresholding Network. Applied Sciences **15**(17) (2026)
2. Li, R., et al.: A Semi-Supervised Framework for Breast Ultrasound Segmentation with Training-Free Pseudo-Label Generation and Label Refinement. arXiv preprint arXiv:2603.06167 (2026)
3. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023)
4. Ma, J., et al.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
5. Zhang, C., et al.: A survey on segment anything model (SAM): Vision foundation model meets medical image analysis. arXiv preprint arXiv:2306.14039 (2023)
6. Song, S., et al.: Structure and Progress Aware Diffusion for Medical Image Segmentation. arXiv preprint arXiv:2603.07889 (2026)
7. Pham, C.M., et al.: Unleashing SAM for Few-Shot Medical Image Segmentation with Dual-Encoder and Automated Prompting. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) (2025)
8. Cao, W., Zhu, K., Zhu, K., Zhang, S., Mao, Y.: WENet: a lightweight dermoscopic image segmentation network with wide edge assistance generated by morphological erosion. Quantitative Imaging in Medicine and Surgery **16**(4) (2026)
9. Yuan, L., Song, J., Fan, Y.: MCNMF-Unet: a mixture Conv-MLP network with multi-scale features fusion Unet for medical image segmentation. PeerJ Computer Science **10**, e1798 (2024)
10. Liang, P., et al.: Vision Foundation Models in Medical Image Analysis: Advances and Challenges. arXiv preprint arXiv:2502.14584 (2025)
11. Xu, C., et al.: Boundary-Aware Test-Time Adaptation for Zero-Shot Medical Image Segmentation. arXiv preprint arXiv:2512.04520 (2025)

12. Shi, P., et al.: Centerline Boundary Dice Loss for Vascular Segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI) (2024)
13. Cheng, J., Ye, J., Deng, Z., et al.: SAM-Med2D. arXiv preprint arXiv:2308.16184 (2023)
14. Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations (ICLR) (2022)
15. Wu, J., Wu, Y., Xie, Y., et al.: SAM-aware Test-time Adaptation for Universal Medical Image Segmentation. arXiv preprint arXiv:2506.05221 (2025)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
17. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ben Ayed, I.: Boundary loss for highly unbalanced segmentation. In: Medical Imaging with Deep Learning (MIDL), pp. 285–296 (2019)
18. Esposito, S., Rebain, D., Onken, A., Li, C., Mac Aodha, O.: VesselSDF: Distance field priors for vascular network reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2025)
19. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L.: ET-Net: A generic edge-attention guidance network for medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 442–450 (2019)
20. Sobel, I.: An Isotropic 3x3 Image Gradient Operator. In: Machine Vision for Three-Dimensional Scenes, Freeman, H. (ed.), pp. 376–379. Academic Press (1990)
21. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, London (1982)