electronics

MDPI

Article

Enhancing Boundary Precision and Long-Range Dependency Modeling in Medical Imaging via Unified Attention Framework

Yi Zhu [1,2], Yawen Zhu [2,3], Hongtao Ma [2], Bin Li [2], Luyao Xiao [2], Xiaxu Wu [2] and Manzhou Li [2,*]

[1] The Second Affiliated Hospital of Nanjing Medical University, Nanjing 210003, China
[2] National School of Development, Peking University, Beijing 100871, China
[3] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
* Correspondence: limanzhou_pku@163.com

Abstract

This study addresses the common challenges in medical image segmentation and recognition, including boundary ambiguity, scale variation, and the difficulty of modeling long-range dependencies, by proposing a unified framework based on a hierarchical attention mechanism. The framework consists of a local detail attention module, a global context attention module, and a cross-scale consistency constraint module, which collectively enable adaptive weighting and collaborative optimization across different feature levels, thereby achieving a balance between detail preservation and global modeling. The framework was systematically validated on multiple public datasets, and the results demonstrated that the proposed method achieved Dice, IoU, Precision, Recall, and F1 scores of 0.886, 0.781, 0.898, 0.875, and 0.886, respectively, on the combined dataset, outperforming traditional models such as U-Net, Mask R-CNN, DeepLabV3+, SegNet, and TransUNet. On the BraTS dataset, the proposed method achieved a Dice score of 0.922, Precision of 0.930, and Recall of 0.915, exhibiting superior boundary modeling capability in complex brain MRI images. On the LIDC-IDRI dataset, the Dice score and Recall were improved from 0.751 and 0.732 to 0.822 and 0.807, respectively, effectively reducing the missed detection rate of small nodules compared to traditional convolutional models. On the ISIC dermoscopy dataset, the proposed framework achieved a Dice score of 0.914 and a Precision of 0.922, significantly improving the accuracy of skin lesion recognition. The ablation study further revealed that local detail attention significantly enhanced boundary and texture modeling, global context attention strengthened long-range dependency capture, and cross-scale consistency constraints ensured the stability and coherence of prediction results. From a medical economics perspective, the proposed framework has the potential to reduce diagnostic costs and improve healthcare efficiency by enabling faster and more accurate image-based clinical decision-making. In summary, the hierarchical attention mechanism presented in this work not only provides an innovative breakthrough in mathematical modeling but also demonstrates outstanding performance and generalization ability in experiments, offering new perspectives and technical pathways for intelligent segmentation and recognition in medical imaging.

Keywords: medical image segmentation; global context modeling; transformer-based hybrid framework; clinical decision support systems; deep learning for lesion detection

check for updates

Academic Editor: Luca Mesin

Received: 14 October 2025
Revised: 2 November 2025
Accepted: 3 November 2025
Published: 5 November 2025

Citation: Zhu, Y.; Zhu, Y.; Ma, H.; Li, B.; Xiao, L.; Wu, X.; Li, M. Enhancing Boundary Precision and Long-Range Dependency Modeling in Medical Imaging via Unified Attention Framework. *Electronics* **2025**, *14*, 4335. https://doi.org/10.3390/electronics14214335

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Medical image segmentation and recognition play a pivotal role in clinical diagnosis and therapeutic decision-making [1]. By accurately delineating lesion regions, clinicians can

intuitively grasp the extent, location, and morphological characteristics of abnormalities, thereby supporting early detection, disease stratification, and treatment evaluation [2]. Medical imaging has become a core tool of modern medicine, widely applied in tumor detection, neurological disease assessment, cardiovascular anomaly analysis, and dermatological lesion recognition [3]. The outcomes of such analyses not only affect diagnostic accuracy and prognostic evaluation but also directly influence personalized treatment planning, radiotherapy protocols, and surgical navigation [4]. Conventional image interpretation depends heavily on physician expertise and subjective judgment, which is both time-consuming and prone to inter-observer variability, making it insufficient to meet the rapidly increasing demand for medical image data analysis [5]. Against this background, the development of efficient, stable, and clinically applicable automated imaging techniques has become particularly urgent [6]. With the widespread adoption of multimodal imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and dermoscopy, the scale and complexity of imaging data continue to increase, further intensifying the demand for automated segmentation and recognition [7].

Such techniques not only alleviate the burden on physicians but also provide objective and reproducible quantitative indices, offering critical support for disease screening, precision treatment, and therapeutic monitoring [8]. From a medical economics standpoint, the advancement of automated and high-precision image analysis methods can substantially reduce diagnostic time, lower the reliance on costly manual interpretation, and optimize resource allocation within healthcare systems, thereby contributing to overall cost-effectiveness and improved patient access to timely diagnosis. Consequently, achieving high-precision lesion segmentation and recognition in complex and heterogeneous imaging environments has emerged as a crucial research direction and challenge in medical artificial intelligence.

Traditional image processing methods, including thresholding, region growing, and graph cuts, achieved moderate success but are sensitive to noise, initialization, and image quality, leading to frequent over-segmentation and limited clinical applicability [9,10]. Deep learning methods, particularly convolutional neural networks (CNNs) and U-Net variants, have advanced automated segmentation by leveraging encoder–decoder structures and skip connections [11,12]. Yet, challenges persist in small-scale lesion recognition due to feature attenuation during down-sampling [13], prompting innovations such as residual connections, atrous convolutions, and multi-scale strategies [14,15]. These methods improved boundary delineation but still struggle to balance fine detail preservation with global consistency in complex images.

Attention mechanisms and Transformer architectures have further enriched segmentation research. Channel attention, such as Squeeze-and-Excitation (SE), enhances lesion-relevant features [16], spatial attention suppresses background noise [17], and self-attention captures long-range dependencies [18]. Transformer-based models like TransUNet and Swin-UNet effectively model global context [19] but often miss fine structural details, especially for blurred or small lesions. Multi-scale fusion approaches, such as pyramid pooling modules (PPM) [20], feature pyramid networks (FPN) [21], and multi-branch structures [22], have improved adaptability to varying lesion sizes [23], though most rely on static stacking or weighted fusion, lacking dynamic consistency modeling across scales.

Despite these advances, challenges remain: ambiguous lesion boundaries due to noise and low contrast [24], feature loss in small-scale lesions [25], and global models overlooking local details, leading to over- or mis-segmentation [26]. To address this, we propose a hierarchical attention–based fine-grained segmentation and recognition framework. It dynamically integrates local detail and global contextual attention, employs a feature enhancement module for small lesions, and introduces cross-scale consistency constraints,

ensuring boundary clarity, contextual completeness, and robust performance in multi-scale scenarios. The major contributions of this work can be summarized as follows:

1. A hierarchical attention mechanism is introduced to achieve adaptive fusion of local and global features;
2. A fine-grained feature enhancement module is designed to improve the detection and segmentation of small-scale lesions;
3. A cross-scale consistency constraint is proposed to guarantee coordination and robustness of segmentation results across scales.

The rest of this paper is organized as follows: Section 2 reviews the research work on medical image segmentation, boundary refinement, and attention-based mechanisms; Section 3 details the dataset and preprocessing pipeline used, and explains the proposed unified attention framework, including local detail attention, global context attention, and cross-scale consistency modules, as well as the specific experimental settings; Section 4 presents quantitative and qualitative experimental results, compares them with the current state-of-the-art methods and performs ablation analysis, discusses the limitations of the model, its potential clinical application value, and future research directions; Section 5 summarizes the work of the entire paper.

2. Related Work

2.1. Medical Image Segmentation Methods

Recent advances in medical image segmentation have been driven by both CNN and Transformer architectures. The U-Net [27] remains the foundational model, combining encoder–decoder pathways and skip connections to fuse semantic and spatial information, achieving strong results in tasks such as brain tumor and cell segmentation. Subsequent variants such as Attention Res-UNet [28] and 3D U-Net [29] enhanced performance via attention gates and volumetric modeling, but challenges remain in handling blurred boundaries and small lesions. To improve feature discrimination, channel attention modules like Squeeze-and-Excitation (SE) [30] and Efficient Channel Attention (ECA) [31] have been incorporated after convolutional blocks to reweight informative channels with minimal computational cost (typically <2% FLOPs). These mechanisms strengthen responses to subtle lesion cues and boundary textures. Beyond channel modeling, non-local and self-attention mechanisms capture long-range dependencies, addressing the limited receptive field of CNNs. Hybrid CNN–Transformer frameworks further extend these ideas. TransUNet [32] integrates Transformer layers into a U-Net backbone, leveraging global attention for better contextual understanding but at higher computational expense. Hierarchical and window-based Transformers, such as Swin-UNet [33], UNETR [34], and UTNet [35], reduce attention complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ via shifted local windows, improving scalability while maintaining global awareness. These modules are typically applied at deeper stages (F_3–F_4) to enhance recognition of small or low-contrast lesions.

Empirical studies show that channel–spatial hybrid attention consistently improves Dice and boundary metrics (by 1–2%) over baseline CNNs, especially for small objects. Lightweight attention modules yield substantial accuracy gains with negligible cost, whereas Transformer-based extensions introduce 20–30% more parameters but deliver stronger generalization across modalities. Overall, these developments reflect a shift toward architectures that balance local precision and global context, improving both robustness and fine-scale segmentation accuracy.

2.2. Applications of Attention Mechanisms in Medical Imaging

With the evolution of deep learning, attention mechanisms have been increasingly applied to medical image analysis to enhance feature selection and contextual modeling [36].

Channel attention, exemplified by the SE module [30], recalibrates channel weights to emphasize lesion-relevant features, and its integration into U-Net or ResNet improved small object recognition in pulmonary nodule and liver lesion segmentation without sacrificing overall accuracy [37,38]. Spatial attention highlights lesion regions within complex backgrounds, as shown by Wu et al. in skin lesion segmentation [39]. More recently, self-attention and Vision Transformer-based methods have enabled long-range dependency modeling, with TransUNet outperforming U-Net in organ and tumor segmentation [32], and ScUNet++ reducing computational cost via hierarchical window-based attention while achieving strong results across datasets [40]. Despite these advances, attention modules may overemphasize certain features and overlook others, while Transformer-based methods often lose local boundary and texture details, leading to segmentation errors in small or blurred lesions. Furthermore, the added computational burden of attention mechanisms remains a challenge for clinical deployment. From the perspective of medical economics, optimizing model efficiency and reducing computational costs are not only technical priorities but also essential factors influencing the scalability and affordability of intelligent diagnostic systems. Efficient attention-based frameworks can lower the operational expenses associated with image analysis, shorten diagnostic cycles, and help allocate limited medical resources more effectively, thus promoting sustainable integration of AI technologies into routine healthcare practice.

2.3. Hierarchical and Multi-Scale Feature Fusion Methods

Lesions in medical imaging often vary greatly in scale and morphology, posing challenges for feature representation [41]. To address this, multi-scale and hierarchical strategies have been developed. Early methods such as the PPM [42] and ASPP in the DeepLab series [43] improved global context modeling but remained limited in boundary refinement and small-lesion capture. FPN further integrated semantic and spatial features via top-down fusion [44], while extensions like U-Net++ and UNet3+ enhanced encoder–decoder interactions with dense skip connections and multi-level decoding, achieving improved performance in complex tasks. HRNet maintained multi-resolution branches to preserve fine details and achieved promising results in vessel and skin lesion segmentation [45], though at high computational cost and with limited cross-scale coordination. Despite these advances, most methods rely on static feature concatenation or weighting [46], lacking dynamic cross-scale modeling and often trading off global context against boundary detail [47]. These limitations highlight the need for more effective dynamic integration of local and global features across scales in medical image segmentation.

3. Materials and Method

3.1. Data Collection

The collection and construction of medical imaging datasets served as a fundamental basis for this study. To ensure the effectiveness and generalization ability of the proposed framework across multi-modal, multi-disease, and multi-scale scenarios, both publicly available and self-collected medical imaging datasets were employed, as shown in Table 1 and Figure 1. The public datasets included the Brain Tumor Segmentation Challenge dataset (BraTS) [48], the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [49], and the International Skin Imaging Collaboration (ISIC) dataset [50]. The BraTS dataset comprises multi-center (MRI) scans collected between 2015 and 2021. It contains four imaging modalities (T1, T1CE, T2, and FLAIR), and each case is accompanied by expert annotations of tumor regions. The spatial resolution is standardized to $240 \times 240 \times 155$, and the primary characteristic lies in its ability to capture multi-modal structural information of brain tumors. The LIDC-IDRI dataset was collected by multiple

institutions in North America from 2004 to 2008, primarily consisting of low-dose helical CT scans of the chest. Each case was independently annotated by four radiologists, featuring substantial variation in nodule diameters, blurred boundaries, and diverse morphologies, making it well-suited for small-scale lesion recognition and segmentation. The ISIC dataset includes dermoscopic images collected across Europe and North America between 2016 and 2020. Images were obtained using high-resolution digital dermatoscopes and cover a wide range of conditions such as melanoma, keratosis, and benign nevi. The resolution ranges from 600 × 450 to 1024 × 768, and the dataset is characterized by pronounced color variations and complex backgrounds. This study was approved by the Biomedical Research Ethics Committee (Approval number IRB00001052-25150, dated 23 January 2025).

Table 1. Summary of medical imaging datasets used in this study.

Dataset	Modality	Collection Time	Resolution/Thickness	Number of Cases
BraTS	MRI (T1, T1CE, T2, FLAIR)	2015–2021	240 × 240 × 155	1500
LIDC-IDRI	CT	2004–2008	1 mm slice thickness	1018
ISIC	Dermoscopy	2016–2020	600 × 450–1024 × 768	2500

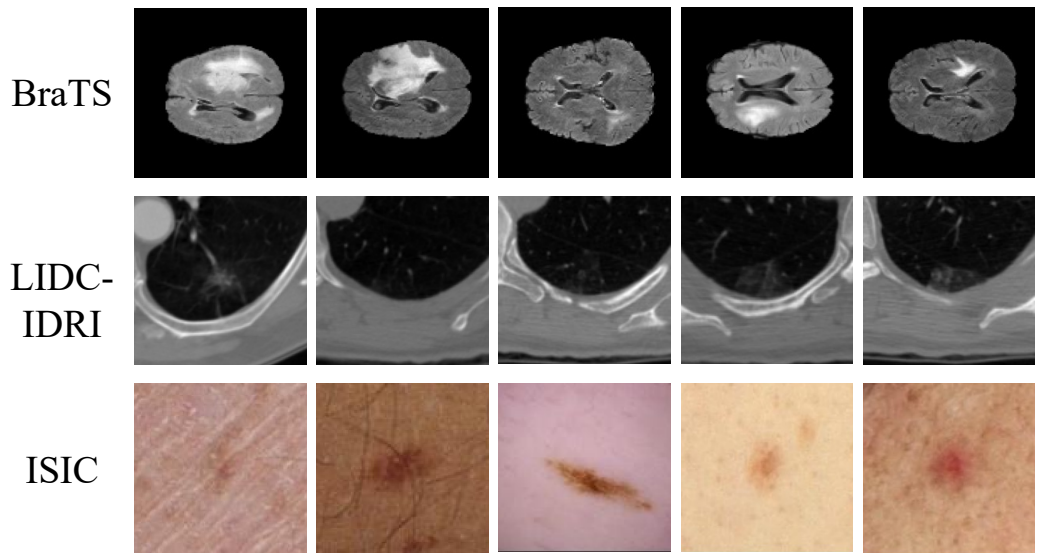


Figure 1. Samples of medical imaging datasets used in this study.

In addition to the public datasets, supplementary self-collected medical imaging data were acquired to fulfill specific experimental requirements. These data were gathered between 2022 and 2023 from two tertiary hospitals in China. The collection included 500 chest CT scans, 300 brain MRI scans, and 200 dermoscopic images. CT data were obtained using a Siemens SOMATOM Definition Flash dual-source CT scanner (Siemens Healthineers, Erlangen, Germany), with a slice thickness of 1 mm, tube voltage of 120 kVp, and coverage from lung apex to diaphragm. A subset of cases included longitudinal follow-up scans, enabling analysis of nodule evolution. MRI data were collected using a GE Discovery MR750 scanner (GE Healthcare, Chicago, IL, USA), including T1, T2, and contrast-enhanced sequences, with isotropic resolution of 1 × 1 × 1 mm. These scans were conducted during hospitalization and were reviewed and annotated by experienced radiologists. Dermoscopic data were acquired using FotoFinder (FotoFinder Systems GmbH, Bad Birnbach, Germany) equipment under controlled illumination and fixed imaging distance, with a resolution of 1024 × 768. The annotations were cross-validated by two dermatologists to ensure accuracy and consistency. The self-collected datasets exhibit strong clinical diversity and

high annotation precision, thereby providing solid support for validating the robustness and applicability of the proposed method in real clinical environments.

3.2. Data Preprocessing and Augmentation

All datasets underwent a unified preprocessing and augmentation pipeline designed to ensure structural consistency across modalities and stable convergence during training. The overall process followed the sequence: spatial standardization (resizing) $\rightarrow$ intensity normalization $\rightarrow$ contrast enhancement $\rightarrow$ pseudo-color mapping (for grayscale modalities) $\rightarrow$ data augmentation. First, all medical images were resized to a fixed spatial resolution of 256×256 pixels to standardize voxel or pixel spacing and facilitate batch training across heterogeneous datasets. Then, intensity normalization was applied to mitigate scanner- and modality-dependent variations, formulated as

$$I_{\text{norm}}(x) = \frac{I(x) - \mu}{\sigma}, \tag{1}$$

where μ and σ denote the mean and standard deviation of the image, respectively. This operation standardizes the gray value range across different images, allowing the model to focus on structural information rather than acquisition artifacts. Since most backbone networks (e.g., U-Net, Swin-UNet, TransUNet) are pre-trained on three-channel natural images, grayscale medical scans were converted into pseudo-RGB form for architectural compatibility. we introduced a monotonic pseudo-color mapping that preserves intensity order while enhancing perceptual separability. For a gray value $g \in [0, 255]$, the mapping function is defined as

$$C(g) = (R(g), G(g), B(g)), \tag{2}$$

where $R(g)$, $G(g)$, and $B(g)$ represent the corresponding red, green, and blue channel intensities. Each component function is defined as a continuous and monotonic transformation of the normalized gray value $\hat{g} = g/255$, i.e.,

$$R(g) = \alpha_R \, \hat{g}^{p_R}, \quad G(g) = \alpha_G \, \hat{g}^{p_G}, \quad B(g) = \alpha_B \, \hat{g}^{p_B}, \tag{3}$$

where $\alpha_R, \alpha_G, \alpha_B \in [0, 1]$ are scaling coefficients controlling color intensity, and $p_R, p_G, p_B > 0$ are channel-wise exponents determining color progression. This formulation ensures that as gray level increases, (R, G, B) vary smoothly and monotonically, maintaining a one-to-one correspondence with the original intensity order:

$$\frac{dR}{dg} > 0, \quad \frac{dG}{dg} > 0, \quad \frac{dB}{dg} > 0, \quad \forall g \in [0, 255].$$

A typical configuration satisfying perceptual uniformity is

$$(\alpha_R, \alpha_G, \alpha_B) = (0.9, 0.8, 1.0), \quad (p_R, p_G, p_B) = (1.2, 1.0, 0.8),$$

which provides a smooth color gradient from dark violet to bright yellow, similar to perceptually uniform colormaps such as viridis. This ensures that the brightness ordering of tissues is preserved while improving the perceptual distinction of structural boundaries. To further enhance global contrast, histogram equalization was applied to the normalized grayscale image before color mapping. The transformation is expressed as

$$g' = \text{round}\left(\frac{L-1}{N} \sum_{i=0}^{g} n_i \right), \tag{4}$$

where L represents the total number of gray levels, N the number of pixels, and n_i the number of pixels with gray value i. This transformation redistributes the gray-level histogram toward a uniform distribution, thereby improving the overall contrast and localizing low-contrast lesions more effectively.

In terms of augmentation, multiple strategies were employed to enhance robustness and generalization. Geometric transformations, including random cropping, rotation, and translation, were introduced to increase data diversity without altering lesion semantics. Specifically, cropping ratios were randomly sampled from the range $[0.8, 1.0]$, rotation angles θ were uniformly drawn from $[-15^\circ, +15^\circ]$, and translations were limited to $\pm 10\%$ of the image size. For image coordinates (x, y), rotation is expressed as

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{5}$$

where θ denotes the rotation angle. By varying θ, the augmentation simulates slight patient pose variations during acquisition. In addition to geometric transformations, color jittering was employed to account for illumination variability and acquisition inconsistencies across imaging devices. The operation randomly perturbs image brightness, contrast, saturation, and hue within controlled ranges while maintaining structural fidelity:

$$\tilde{I}(x) = \mathrm{Hue}\big(\mathrm{Sat}(\beta \cdot I(x) + \delta_b)\big), \tag{6}$$

where $\mathrm{Sat}(\cdot)$ denotes a linear interpolation between the original image and its grayscale version to control color intensity, and $\mathrm{Hue}(\cdot)$ applies a cyclic shift to the hue component in HSV color space, $\beta \sim \mathcal{U}(0.8, 1.2)$ controls contrast and $\delta_b \sim \mathcal{U}(-0.2, 0.2)$ adjusts brightness. These perturbations effectively simulate variations in illumination and sensor response without altering anatomical semantics, thereby enhancing the model's robustness to photometric variations.

$$\tilde{X} = \lambda X_i + (1-\lambda) X_j, \quad \tilde{Y} = \lambda Y_i + (1-\lambda) Y_j, \tag{7}$$

where $\lambda \in [0, 1]$ follows a Beta distribution $\mathrm{Beta}(\alpha, \alpha)$ with $\alpha = 0.4$. This setting balances between preserving structural integrity and introducing controlled interpolation, effectively mitigating overfitting and enhancing generalization to unseen variations. CutMix, on the other hand, replaces a randomly cropped region of one image with a corresponding region from another image, formulated as

$$\tilde{X} = M \odot X_i + (1 - M) \odot X_j, \quad \tilde{Y} = \lambda Y_i + (1-\lambda) Y_j, \tag{8}$$

where M denotes a binary mask matrix and $\odot$ indicates element-wise multiplication. The cropping region was randomly chosen to occupy 20–40% of the total area, and λ was drawn from the same $\mathrm{Beta}(1.0, 1.0)$ distribution as recommended in prior studies. This strategy strengthens the model's capability in handling incomplete or partially occluded information, thereby improving robustness in segmenting blurred boundaries and small-scale lesions. Through the integration of these preprocessing and augmentation approaches, input data were fully standardized and diversified, ensuring stable training dynamics and enhanced segmentation performance across modalities.

3.3. Proposed Method

3.3.1. Overall

The overall method is built upon a unified dual-branch architecture with a shared encoder–decoder backbone. The processed medical image x is first fed into a shared encoder E, which is implemented as a hybrid U-Net–Transformer structure consisting

of convolutional stem layers followed by hierarchical transformer blocks. The encoder extracts multi-scale features $\{F_1, F_2, F_3, F_4\}$ through progressive down-sampling, where spatial resolution gradually decreases while semantic abstraction increases. The parameters of E are fully shared between the two downstream branches to ensure semantic consistency and parameter efficiency. Both branches thus operate on identical feature maps produced by the same encoder weights. To facilitate faster convergence and stable representation learning, E is initialized with ImageNet-pretrained weights and then fine-tuned on medical imaging data. The high-resolution layers F_1 and F_2 are directed to the local detail attention branch, which focuses on fine-grained contour and texture modeling. This branch employs channel–spatial joint gating and boundary-sensitive residual units to assign normalized attention weights within $[0, 1]$ to regions containing potential boundaries or microstructures, yielding enhanced local representations $\{L_1, L_2\}$. A texture–edge embedding unit further preserves structural granularity and mitigates noise, while lightweight depthwise separable convolutions and pointwise re-calibration stabilize responses for small-scale lesions. The deeper layers F_3 and F_4 are processed by the global context attention branch, which constructs query, key, and value tensors for self-attention aggregation. Relationships are first captured within non-overlapping local windows and subsequently extended across windows via shifted-window interaction, enabling long-range dependency modeling and the generation of globally enriched features $\{G_3, G_4\}$ containing organ topology and lesion-level semantics. Relative position encoding is incorporated to maintain spatial consistency and reduce boundary mismatches. While the two branches operate independently in their respective resolution spaces, feature interaction is implicitly maintained through the shared encoder and explicitly reinforced during the fusion stage. At the fusion stage, $\{L_1, L_2\}$ and $\{G_3, G_4\}$ are projected to a unified embedding dimension by a dynamic cross-scale fusion module and combined through pixel-wise adaptive weighting and concatenation to yield the fused representation Z. The decoder adopts a U-shaped progressive up-sampling strategy, aligning Z with encoder features via skip connections to produce multi-scale predictions $\{P_1, P_2, P_3, P_4\}$. A cross-scale consistency constraint enforces bidirectional alignment among predictions, stabilizing outputs across scales. The overall loss integrates region-overlap and boundary-similarity terms for joint optimization. A global pooling operation applied to Z, followed by a lightweight classification head, generates lesion class probabilities, enabling joint segmentation and recognition within a unified backbone. During inference, the model outputs multi-scale masks and classification scores from coarse to fine, refined via an adaptive ensemble guided by scale reliability and uncertainty estimation. Residual connections, layer normalization, and attention temperature scaling ensure numerical stability, while separable convolutions and sparse attention mechanisms effectively control computational cost, forming an integrated "encoding–dual-branch attention–dynamic fusion–consistency constraint–multi-head output" framework for lesion segmentation and recognition. To facilitate understanding of the computational flow and feature resolutions, Table 2 summarizes the main operations and tensor dimensions across the proposed framework.

Table 2. An overview of major operations and corresponding tensor dimensions throughout the proposed architecture. H and W denote the input image height and width, and C represents the base channel width.

Stage	Operation/Module	Input Tensor	Output Tensor
Input	Preprocessed image	$I \in \mathbb{R}^{3\times H\times W}$	—
Encoder E	Conv + Patch embedding (Stage 1)	I	$F_1 \in \mathbb{R}^{C\times \frac{H}{2}\times \frac{W}{2}}$
	Hierarchical down-sampling + Transformer blocks (Stage 2–4)	F_1	$\{F_2, F_3, F_4\}$, where $F_2 \in \mathbb{R}^{2C\times \frac{H}{4}\times \frac{W}{4}}$, $F_3 \in \mathbb{R}^{4C\times \frac{H}{8}\times \frac{W}{8}}$, $F_4 \in \mathbb{R}^{8C\times \frac{H}{16}\times \frac{W}{16}}$
Local Detail Attention Branch	Channel–spatial joint gating + boundary-sensitive residuals	$\{F_1, F_2\}$	$\{L_1, L_2\}$ (same shape as inputs)
Global Context Attention Branch	Windowed self-attention + shifted windows + position encoding	$\{F_3, F_4\}$	$\{G_3, G_4\}$ (same shape as inputs)
Dynamic Cross-Scale Fusion	Projection + adaptive pixel-wise weighting + concatenation	$\{L_1, L_2, G_3, G_4\}$	$Z \in \mathbb{R}^{4C\times \frac{H}{4}\times \frac{W}{4}}$
Decoder	Progressive up-sampling + skip connections	Z	$\{P_1, P_2, P_3, P_4\}$ (multi-scale predictions)
Cross-Scale Consistency	Bidirectional alignment among $\{P_i\}$	$\{P_1, P_2, P_3, P_4\}$	Consistent feature maps
Classification Head	Global pooling + fully connected layer	Z	Class probabilities $p \in \mathbb{R}^K$
Output	Segmentation masks + lesion class probabilities	—	$\{\hat{Y}, p\}$

3.3.2. Local Detail Attention Module

The local detail attention module undertakes the core task of fine-grained modeling for small-scale lesions and ambiguous boundary regions. Its fundamental idea is to employ a joint modeling mechanism of spatial attention and channel attention, enabling the network to adaptively highlight lesion-related boundary and texture features while suppressing background noise and redundant information. As shown in Figure 2, this module first receives the multi-scale feature maps output from the backbone encoder. Let the input feature be $F \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of channels, and H and W represent the height and width of the feature map. To enhance sensitivity to small-scale lesions, the design applies global average pooling and global max pooling along the channel dimension to obtain two sets of channel descriptors, which are then transformed through shared fully connected layers with nonlinear activation, followed by a sigmoid function to generate channel-wise weights w_c. The channel-recalibrated feature is thus expressed as $F_c = w_c \odot F$, allowing different channels to be reweighted based on their importance to lesion characterization, thereby emphasizing structural edges and fine textures. Simultaneously, the spatial attention branch aggregates spatial descriptors through max pooling and average pooling across the channel dimension, which are then passed through a convolutional layer and a sigmoid mapping to form a spatial weight map $M_s \in \mathbb{R}^{H\times W}$. Applying this to the input produces $F_s = M_s \odot F$, reinforcing the discriminative capacity of the network at lesion boundaries. Finally, the outputs of channel and spatial attention are fused as $F' = F_c + F_s$, yielding an enhanced feature representation that jointly models both channel and spatial dimensions.

In terms of architectural parameters, a 1×1 convolution is employed to compress channel dimensions and reduce computational cost, while residual connections are introduced to ensure stable gradient propagation. Each fully connected layer in the attention branches uses $\frac{C}{r}$ hidden units with a reduction ratio $r = 16$, striking a balance between representational power and parameter efficiency. The spatial attention convolution kernel is set to 7×7, allowing the capture of broader spatial dependencies while maintaining

computational feasibility, and better simulating the transitional characteristics of lesion boundaries. To further strengthen sensitivity to small lesions, a boundary-aware loss is integrated into training, compelling the module to achieve finer resolution at lesion edges by penalizing discrepancies between predicted and annotated boundaries.

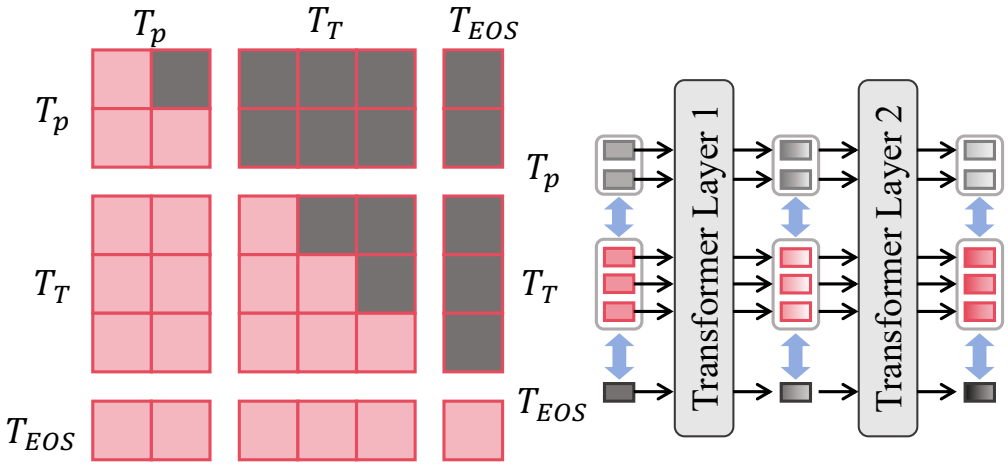


Figure 2. This illustrates the architecture of the Local Detail Attention Module.

From a mathematical perspective, this module can be regarded as adaptive reparameterization of feature distributions by jointly optimizing channel and spatial weights. Channel attention acts as a learnable weighted inner product in feature space, highlighting the most discriminative subspaces for segmentation, while spatial attention functions as a position-dependent weighting function that amplifies gradient variations around lesion boundaries. Through this dual-weighting mechanism, the model maintains high discriminative ability even in scenarios involving blurred boundaries and sparse small lesions. Empirical results demonstrate that this module significantly improves Dice and intersection over union (IoU) metrics, particularly excelling in cases involving micro-nodules and dermatological lesions, thereby validating its design rationale and its effectiveness in fine-grained medical image segmentation.

3.3.3. Global Context Attention Module

The global context attention module operates on mid-to-high-level features extracted from the encoder, as shown in Figure 3. Let $F_3 \in \mathbb{R}^{C_3 \times H_3 \times W_3}$ and $F_4 \in \mathbb{R}^{C_4 \times H_4 \times W_4}$ represent two levels of features with $C_3 = 256, H_3 = W_3 = 64$ and $C_4 = 512, H_4 = W_4 = 32$, respectively. The module is composed of two hierarchical global modeling stages, each consisting of two window-based self-attention layers and two convolutional feed-forward layers, yielding a total of four layers. In the F_3 branch, non-overlapping 8×8 windows with $h_3 = 8$ heads and per-head dimension $d_3 = 32$ are applied. In the F_4 branch, shifted 8×8 windows are employed with $h_4 = 12$ heads and per-head dimension $d_4 = 32$. To avoid duplicating the standard Softmax attention formulation, a kernelized linear attention mechanism with explicit relative positional bias and learnable gating is introduced.

Specifically, projections are computed as $Q = W_q X$, $K = W_k X$, $V = W_v X$ through 1×1 convolutions, followed by linearized aggregation using a positive mapping $\phi(\cdot)$, where $\phi(z) = \mathrm{ELU}(z) + 1$. The output Y is defined as

$$Y = D^{-1}\Big(\phi(Q)\big(\phi(K)^\top V\big)\Big), \quad D = \mathrm{diag}\Big(\phi(Q)\big(\phi(K)^\top \mathbf{1}\big)\Big), \tag{9}$$

with head-wise modulation by relative positional terms B and learnable gates $g \in (0, 1)$:

$$\widetilde{Y} = \sum_{m=1}^{h} g_m \Big(Y_m + \big(\phi(Q_m) \odot P_m\big)\big(\phi(K_m)^\top \mathbf{b}_m\big) \Big) + B, \tag{10}$$

where $\odot$ denotes the Hadamard product, P_m represents positional encoding, and $\mathbf{b}_m$ is a learnable bias vector. By the associativity of matrix multiplication, the above linear attention admits a low-rank reformulation. First, $S = \phi(K)^\top V \in \mathbb{R}^{d\times C}$ and $z = \phi(K)^\top \mathbf{1} \in \mathbb{R}^d$ are computed, allowing Y and D to be obtained with $O(Nd + Cd)$ complexity. Compared to quadratic-cost global attention, this formulation substantially reduces memory and computation when $N = HW$ is large. The proof is based on the decomposition

$$\phi(Q)\big(\phi(K)^\top V\big) = \sum_{i=1}^{N} \phi(Q_{:i})\big(\phi(K_{:i})^\top V_{i:}\big), \tag{11}$$

which directly follows from column-wise expansion of matrix multiplication, thus transforming $N \times N$ interactions into products involving $N \times d$ and $d \times C$. This ensures scalability and stable gradient propagation with increasing resolution.

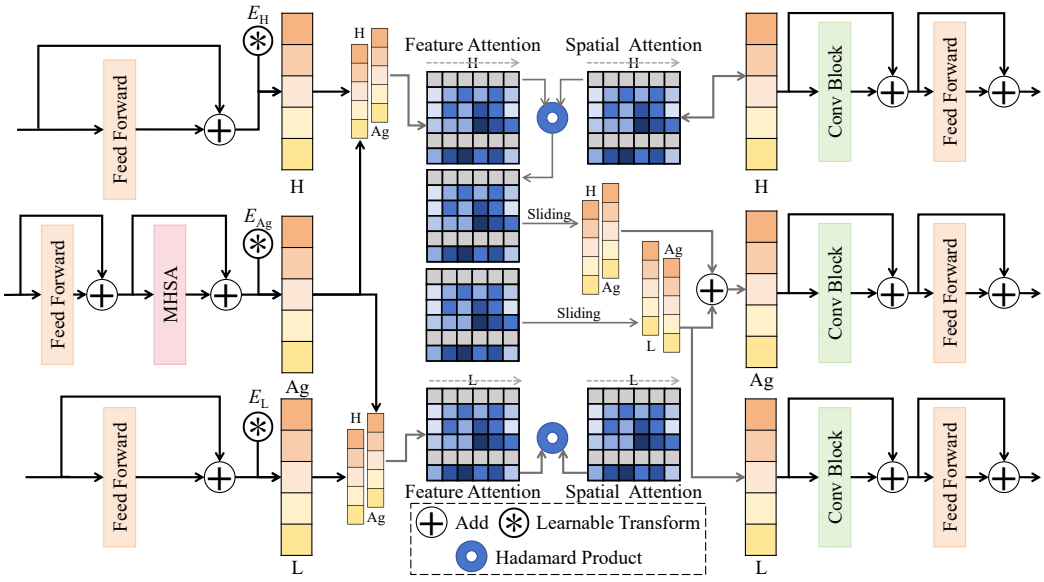


Figure 3. The figure illustrates the architecture of the Global Context Attention Module.

The feed-forward layers employ depthwise separable 3×3 convolutions followed by 1×1 pointwise convolutions, with a channel expansion ratio of 4. For the F_3 branch, the channels evolve as $256 \to 1024 \to 256$, and for the F_4 branch, as $512 \to 2048 \to 512$. Layer normalization and SiLU activation are applied within each block, while alternating shifted windows restore long-range dependencies across partitions. The two global features are aligned and fused at a 64×64 resolution through upsampling and 1×1 projection, producing $G \in \mathbb{R}^{384\times64\times64}$ (from concatenating the 256 channels of F_3 with 128 channels from the upsampled F_4 and compressing them). This is integrated with the enhanced features $L \in \mathbb{R}^{128\times64\times64}$ from the local detail attention module via gated fusion:

$$Z = \sigma\big(W_g[G; L]\big) \odot G + \big(1 - \sigma\big(W_g[G; L]\big)\big) \odot L, \tag{12}$$

where W_g is a 1×1 convolution and $\sigma(\cdot)$ is a Sigmoid activation. This joint design is mathematically equivalent to a convex combination over feature subspaces. Let the expected segmentation risk $R(\cdot)$ be Lipschitz continuous. Then, there exists $\kappa > 0$ such that

$$R(Z) \le \sigma^\star R(G) + (1 - \sigma^\star) R(L) + \kappa \|G - L\|_2, \tag{13}$$

where $\sigma^\star$ is the optimal gating weight. As global-local consistency improves and $\|G - L\|_2$ diminishes, the upper bound on risk contracts, thereby explaining the stabilizing benefit when jointly trained with cross-scale consistency constraints. Applied to the present task, this linear attention mechanism preserves global dependencies and organ topology while reducing computation from $O(N^2)$ to $O(N)$. Shifted windows and relative positional terms maintain spatial consistency and inter-region connectivity, while gated fusion aligns large-scale semantics with small-scale boundary details at a unified scale. This markedly improves boundary clarity and recall for small lesions and achieves controllable memory consumption and stable training under practical configurations (e.g., 512×512 inputs with $N = 4096$ at 64×64 resolution). The design aligns with the structural diagram in its incorporation of learnable transformations and Hadamard operations, further enhancing inter-layer accessibility and noise robustness.

3.3.4. Cross-Scale Consistency Constraint Module

The cross-scale consistency constraint module is designed to address the inconsistency problem among predictions at different scales, ensuring that segmentation results maintain semantic and boundary coherence across multi-level outputs. As shown in Figure 4, this module operates directly on the multi-scale predictions P_1, P_2, P_3, P_4 generated by the decoder, where $P_1 \in \mathbb{R}^{C\times128\times128}$, $P_2 \in \mathbb{R}^{C\times64\times64}$, $P_3 \in \mathbb{R}^{C\times32\times32}$, $P_4 \in \mathbb{R}^{C\times16\times16}$, and $C = 2$ denotes the binary segmentation mask channels.

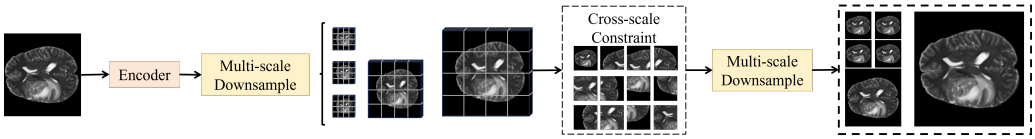


Figure 4. Illustration of the cross-scale consistency constraint module.

The core idea of the module is to establish coordination between local and global predictions across scales through bidirectional mapping and consistency constraints. Specifically, a 1×1 convolution is first applied to project each scale prediction into a unified embedding dimension $d = 64$, followed by upsampling and downsampling operations that align all predictions to an intermediate resolution of 64×64, forming the aligned set $\hat{P}_i$. At this scale, the cross-scale consistency loss is introduced:

$$\mathcal{L}_{cs} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}|\hat{P}_i - \hat{P}_j|_2^2, \tag{14}$$

where $N = 4$ indicates the number of scales. This loss enforces convergence of predictions within the embedding space toward a consistent distribution, thereby mitigating prediction fluctuations and boundary misalignments caused by scale discrepancies. Furthermore, to enhance boundary-level consistency, a boundary consistency term is defined as

$$\mathcal{L}_{bcs} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}|\nabla\hat{P}_i - \nabla\hat{P}_j|_1, \tag{15}$$

where ∇ represents the Sobel operator used to extract boundary gradients of the predicted masks. This constraint ensures that boundaries across different scales preserve similar contour shapes, thereby improving stability in small lesions and blurred boundary regions.

The module consists of two major sublayers: the scale alignment layer and the consistency constraint layer. The scale alignment layer comprises 1×1 convolution and bilinear interpolation, with channels unified to 64, while LayerNorm is applied to high-level features to guarantee numerical stability. The consistency constraint layer employs

a combination of cosine similarity correction and Euclidean distance penalties to prevent prediction collapse into overly smooth distributions. To reduce computational overhead, loss calculation is restricted to a sampled point set S with a sampling ratio of 25%, which can be theoretically regarded as a Monte Carlo approximation of full-pixel consistency. Its unbiasedness guarantees that the expected sampled consistency constraint equals the global constraint, i.e.,

$$\mathbb{E}_S[\mathcal{L}_{cs}(S)] = \mathcal{L}_{cs}, \tag{16}$$

thereby proving that no systematic bias is introduced by sampling. When combined with the local detail attention module and the global context attention module, this module ensures that fine-grained and global semantic enhancements provided by the former are further consolidated by scale-consistent convergence at the output level. The overall optimization objective is formulated as

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\mathcal{L}_{cs} + \beta\mathcal{L}_{bcs}, \tag{17}$$

where $\mathcal{L}_{seg}$ is the primary segmentation loss, and α, β are balance coefficients. Mathematically, this design is equivalent to introducing a convex consistency constraint into the multi-scale prediction space, thereby maintaining topological stability across scales. When applied to medical imaging tasks, the module significantly reduces the drift of small lesion predictions across levels, producing final fused results with sharper boundaries and enhanced cross-domain robustness, with particular improvements in Dice and IoU metrics.

3.4. Experimental Setup

3.4.1. Configurations

To comprehensively evaluate the effectiveness of the proposed method, experiments were conducted on several publicly available medical imaging datasets covering different modalities and task scenarios. Specifically, the BraTS brain tumor segmentation dataset [51], the LIDC-IDRI lung nodule detection dataset [52], and the ISIC skin lesion segmentation dataset [50] were employed. These datasets are representative, encompassing MRI, CT, and dermoscopic images, thereby enabling validation of the adaptability and robustness of the proposed model across diverse medical imaging modalities. For data partitioning, each dataset was divided into training and testing subsets at a ratio of 7:3, with an additional 10% of the training data reserved as a validation set for model selection and hyperparameter tuning. The division was stratified according to lesion presence and size distribution to ensure balanced class proportions and comparable lesion characteristics across subsets, thus preventing sampling bias and ensuring fair evaluation. All input images were first resized to a fixed spatial resolution and then underwent standardized preprocessing procedures as described in Section 3.2. During training, multiple augmentation strategies—including random cropping, rotation, flipping, color jittering, Mixup, and CutMix—were applied to improve robustness and generalization. Model training was conducted on an NVIDIA A100 GPU platform using the PyTorch 2.5.1 deep learning framework. The AdamW optimizer was employed with an initial learning rate of 1×10^{-4}, dynamically adjusted via a cosine annealing schedule. The batch size was set to 16, and training proceeded for up to 200 epochs.

3.4.2. Evaluation Metrics

To thoroughly assess performance in medical image segmentation and recognition tasks, several widely used metrics were applied, including the Dice similarity coefficient (DSC) [53], IoU [54], precision [55], recall [55], and F1-score [56,57]. These metrics reflect

segmentation accuracy, boundary alignment, and sensitivity to small lesion detection from different perspectives. Their mathematical definitions are as follows:

$$\text{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}, \tag{18}$$

$$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}, \tag{19}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{20}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{21}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{22}$$

where P represents the set of pixels predicted as lesions, G denotes the ground-truth lesion set, $|P|$ and $|G|$ indicate the number of pixels in each set, $|P \cap G|$ is the intersection between prediction and ground truth, and $|P \cup G|$ denotes their union. For confusion-matrix-based metrics, TP (true positive) refers to the number of pixels correctly predicted as lesions, FP (false positive) refers to pixels incorrectly predicted as lesions, and FN (false negative) refers to pixels incorrectly predicted as non-lesions. The Dice coefficient and IoU measure the overlap between predictions and ground truth, precision and recall reflect prediction accuracy and completeness, and the F1-score balances the performance of precision and recall through harmonic mean. Together, these metrics provide a comprehensive and objective evaluation of segmentation and recognition performance.

3.4.3. Baseline Models

To verify the effectiveness and advantages of the proposed method, several representative deep learning segmentation models were selected as baselines, including U-Net [27], SegNet [58], DeepLabV3+ [43], Mask R-CNN [59], Attention-UNet [60], Swin-UNet [33], TransUNet [32], SimpleUNet [61] and MedDINOv3 [62]. These models are influential in medical image segmentation and reflect the performance levels of mainstream approaches.

4. Results and Discussion

4.1. Overall Performance of Different Methods Across All Datasets

The purpose of this experiment was to provide a comprehensive assessment of the proposed hierarchical attention framework across multiple medical imaging modalities and to compare its performance with those of representative baseline models. All methods were evaluated under a unified dataset partition and metric protocol to ensure the fairness and reproducibility of comparisons. The reported results were derived from four benchmark datasets—BraTS, LIDC-IDRI, ISIC, and the self-collected clinical dataset—covering MRI, CT, dermoscopic, and multimodal imaging scenarios. For each dataset, segmentation performance was quantified using Dice, IoU, Precision, Recall, and F1-score, which were first computed at the image level and subsequently averaged across all test samples. The final values presented in Table 3 correspond to the macro-averaged results across datasets, ensuring that each imaging modality contributed equally regardless of dataset size or class distribution.

As shown in Table 3, early convolution-based models such as U-Net and SegNet provided stable segmentation results, yet their limited receptive fields and lack of explicit contextual modeling restricted their performance, particularly when dealing with irregular or large lesions. Mask R-CNN, though incorporating a region-based detection mecha-

nism, did not yield substantial gains in dense pixel-level segmentation accuracy due to its coarse-grained feature representation. DeepLabV3+ achieved a notable improvement by utilizing atrous convolutions and multi-scale feature aggregation, effectively capturing richer contextual cues. Attention-UNet further enhanced local–global feature discrimination by introducing attention gates, improving Precision and Recall, especially around boundary regions. Building upon this, Swin-UNet leveraged hierarchical Transformer blocks to model long-range dependencies while preserving spatial locality, achieving more balanced results across all metrics. TransUNet continued this trend by integrating Transformer modules into a CNN backbone, further strengthening global semantic coherence and yielding competitive results on Dice and IoU. More recently, SimpleUNet demonstrated the advantage of streamlined encoder–decoder interactions with efficient attention reparameterization, striking a good balance between accuracy and computational cost. MedDINOv3, incorporating self-distillation and vision-language pretraining, achieved superior generalization across multiple datasets by leveraging large-scale medical priors. Finally, the proposed hierarchical attention framework achieved the highest scores across all evaluation metrics, surpassing both convolution-based and Transformer-based baselines. This performance gain confirms the framework's ability to effectively integrate local detail modeling, global context understanding, and cross-scale consistency. The results demonstrate that the proposed architecture not only achieves state-of-the-art accuracy but also maintains robustness and adaptability across different medical imaging modalities.

Table 3. Overall performances of different methods across all datasets. Metrics include Dice, IoU, Precision, Recall, and F1. Higher is better.

Method	Dice	IoU	Precision	Recall	F1
U-Net [27]	0.837	0.725	0.854	0.819	0.837
SegNet [58]	0.819	0.708	0.841	0.799	0.820
DeepLabV3+ [43]	0.853	0.744	0.872	0.835	0.853
Mask R-CNN [59]	0.836	0.723	0.861	0.809	0.836
Attention-UNet [60]	0.861	0.757	0.873	0.848	0.861
Swin-UNet [33]	0.868	0.761	0.877	0.854	0.868
TransUNet [32]	0.872	0.765	0.880	0.858	0.872
SimpleUNet [61]	0.876	0.770	0.886	0.862	0.876
MedDINOv3 [62]	0.879	0.774	0.889	0.867	0.879
Proposed	**0.886**	**0.781**	**0.898**	**0.875**	**0.886**

4.2. Detail Analysis

The purpose of this experiment was to evaluate the segmentation and recognition performances of different models on three representative medical imaging datasets, thereby verifying the generalizability and robustness of the proposed method across different modalities and task scenarios.

As shown in Table 4, the proposed hierarchical attention framework consistently outperformed all baseline models across all datasets and evaluation metrics. On the BraTS dataset, which involves complex tumor morphology and heterogeneous intensity distributions in MRI, the proposed model achieved a Dice score of 0.922 and an IoU of 0.853, surpassing both CNN-based (e.g., U-Net, DeepLabV3+) and Transformer-based methods (e.g., Swin-UNet, TransUNet). This indicates its superior ability to capture both fine-grained boundaries and long-range spatial dependencies. For the LIDC-IDRI dataset, which presents high variability in nodule size and subtle contrast differences in CT imaging, the proposed model achieved the highest Dice score of 0.822, effectively reducing missed detections of small nodules. On the ISIC dataset, which focuses on dermoscopic lesion segmentation, the model obtained a Dice score of 0.914 and an IoU of 0.842, demonstrating

strong adaptability to surface texture variations and illumination inconsistencies. Finally, on the self-collected dataset, comprising diverse real-world clinical data, the proposed method achieved a Dice of 0.889 and IoU of 0.791, confirming its robustness and clinical applicability under non-standardized imaging conditions. In addition to segmentation accuracy, computational efficiency was quantitatively analyzed in terms of memory consumption and inference time, as shown in Figures 5 and 6. The proposed framework exhibited the lowest GPU memory usage (740 MB) and shortest inference time (0.35 s per image) among all evaluated models, highlighting its efficient attention design and suitability for real-time deployment in clinical environments. Furthermore, Figure 7 presents the Receiver Operating Characteristic (ROC) curves of different models across the four datasets. The proposed method consistently achieved the highest area under the curve (AUC) values, reflecting its superior discriminative capability and stability in distinguishing lesion and non-lesion regions.

Overall, these quantitative and qualitative results confirm that the proposed unified attention framework effectively balances local detail refinement, global context modeling, and computational efficiency, achieving state-of-the-art performance across diverse medical imaging modalities and clinical scenarios.

Table 4. Performances of different methods on four datasets: BraTS (MRI), LIDC-IDRI (CT), ISIC (dermoscopy), and the self-collected dataset. Metrics: Dice, IoU, Precision, Recall, and F1.

Dataset	Method	Dice	IoU	Precision	Recall	F1
BraTS	U-Net	0.883	0.789	0.895	0.872	0.883
	SegNet	0.861	0.758	0.880	0.845	0.862
	DeepLabV3+	0.892	0.804	0.908	0.878	0.893
	Mask R-CNN	0.865	0.764	0.901	0.835	0.867
	Attention-UNet	0.889	0.799	0.907	0.874	0.889
	Swin-UNet	0.895	0.811	0.910	0.881	0.895
	TransUNet	0.904	0.823	0.914	0.895	0.904
	SimpleUNet	0.909	0.832	0.919	0.902	0.909
	MedDINOv3	0.915	0.841	0.925	0.906	0.915
	Proposed	**0.922**	**0.853**	**0.930**	**0.915**	**0.922**
LIDC-IDRI	U-Net	0.751	0.606	0.774	0.732	0.752
	SegNet	0.740	0.591	0.768	0.718	0.741
	DeepLabV3+	0.783	0.644	0.806	0.764	0.784
	Mask R-CNN	0.776	0.635	0.812	0.744	0.776
	Attention-UNet	0.789	0.649	0.816	0.759	0.789
	Swin-UNet	0.792	0.654	0.818	0.766	0.792
	TransUNet	0.795	0.660	0.812	0.779	0.795
	SimpleUNet	0.805	0.673	0.828	0.786	0.805
	MedDINOv3	0.814	0.683	0.835	0.794	0.814
	Proposed	**0.822**	**0.697**	**0.840**	**0.807**	**0.822**
ISIC	U-Net	0.876	0.781	0.892	0.862	0.876
	SegNet	0.855	0.748	0.876	0.836	0.855
	DeepLabV3+	0.885	0.790	0.902	0.868	0.885
	Mask R-CNN	0.867	0.770	0.901	0.835	0.867
	Attention-UNet	0.883	0.788	0.904	0.861	0.883
	Swin-UNet	0.889	0.797	0.907	0.872	0.889
	TransUNet	0.898	0.809	0.910	0.886	0.898
	SimpleUNet	0.905	0.821	0.915	0.892	0.905
	MedDINOv3	0.909	0.830	0.918	0.898	0.909
	Proposed	**0.914**	**0.842**	**0.922**	**0.906**	**0.914**

Table 4. *Cont.*

Dataset	Method	Dice	IoU	Precision	Recall	F1
Self-Collected	U-Net	0.792	0.649	0.818	0.775	0.792
	SegNet	0.773	0.627	0.806	0.758	0.773
	DeepLabV3+	0.812	0.674	0.836	0.792	0.812
	Mask R-CNN	0.804	0.662	0.831	0.780	0.804
	Attention-UNet	0.821	0.683	0.842	0.801	0.821
	Swin-UNet	0.829	0.692	0.849	0.808	0.829
	TransUNet	0.835	0.701	0.855	0.816	0.835
	SimpleUNet	0.842	0.713	0.863	0.823	0.842
	MedDINOv3	0.848	0.721	0.870	0.828	0.848
	Proposed	**0.872**	**0.752**	**0.889**	**0.848**	**0.872**

Note: Bold values indicate the best performance within each dataset.

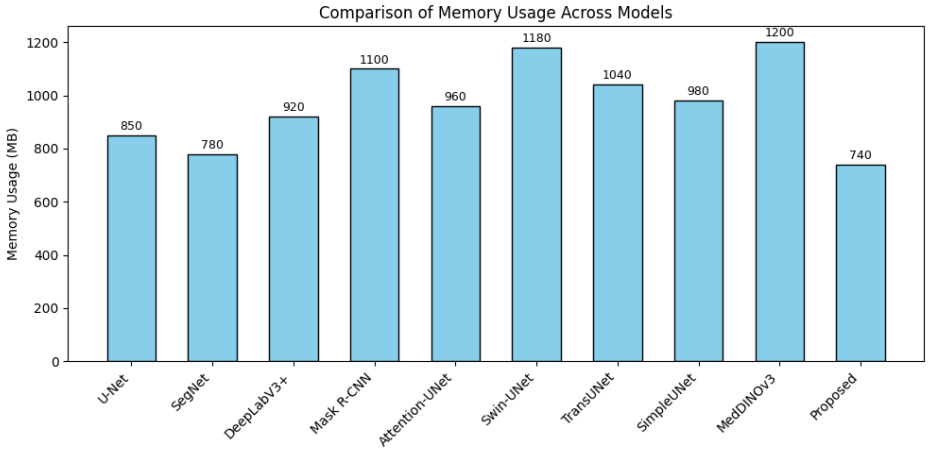


Figure 5. Comparison of memory consumption among different segmentation models.

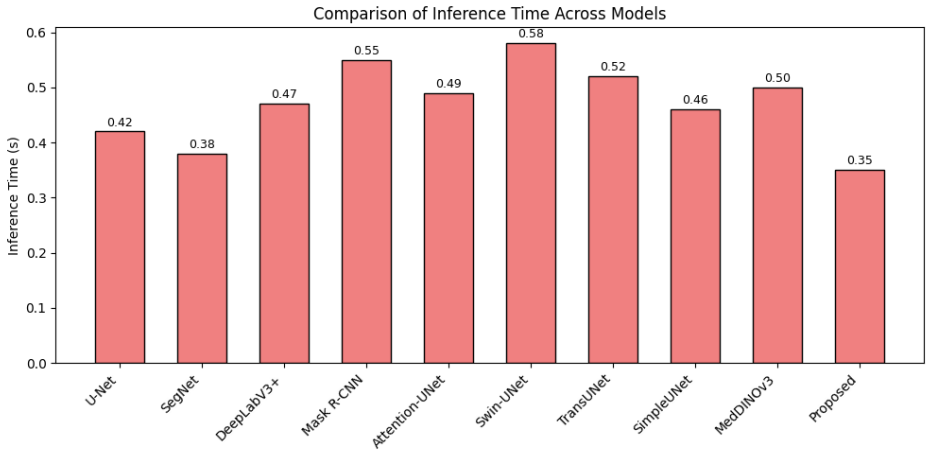


Figure 6. Comparison of inference time per image for various segmentation models.

As shown in Figure 8, qualitative comparisons across MRI, CT, dermoscopic, and ultrasound modalities reveal distinct feature modeling behaviors among the evaluated segmentation models. U-Net and SegNet, both convolution-based encoder–decoder architectures, effectively preserved local spatial details through skip connections but lacked the ability to model long-range dependencies, resulting in coarse or incomplete boundaries. DeepLabV3+ expanded the receptive field via atrous convolutions, corresponding to sparse neighborhood sampling that improved context perception but caused discontinuity in fine-grained structures. Mask R-CNN, while efficient for lesion localization, relied on region proposals

and struggled with irregular or contiguous lesion boundaries. Attention-UNet incorporated spatial gating to enhance focus on salient regions, yet its channel interaction remained limited. Swin-UNet and TransUNet introduced Transformer-based hierarchical self-attention to strengthen global dependency modeling, achieving better semantic consistency but at higher computational cost. SimpleUNet improved architectural compactness with simplified skip pathways, providing a favorable balance between accuracy and efficiency, whereas MedDINOv3 leveraged foundation-model-level pretraining to enhance generalization across modalities but required substantial computational resources. In contrast, the proposed hierarchical attention framework dynamically integrates local boundary refinement and global semantic aggregation, reinforced by cross-scale consistency constraints that stabilize multi-resolution learning. This design yields sharper lesion delineation, higher sensitivity to small-scale structures, and stronger cross-modality generalization, as reflected by consistently superior IoU performance across all datasets.

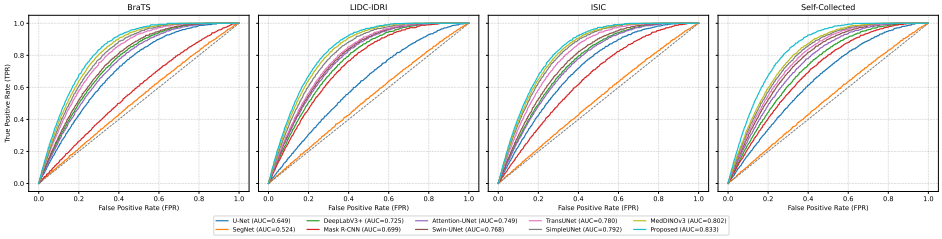


Figure 7. Receiver operating characteristic (ROC) curves of different segmentation models on four medical imaging datasets (BraTS, LIDC-IDRI, ISIC, and self-collected dataset).

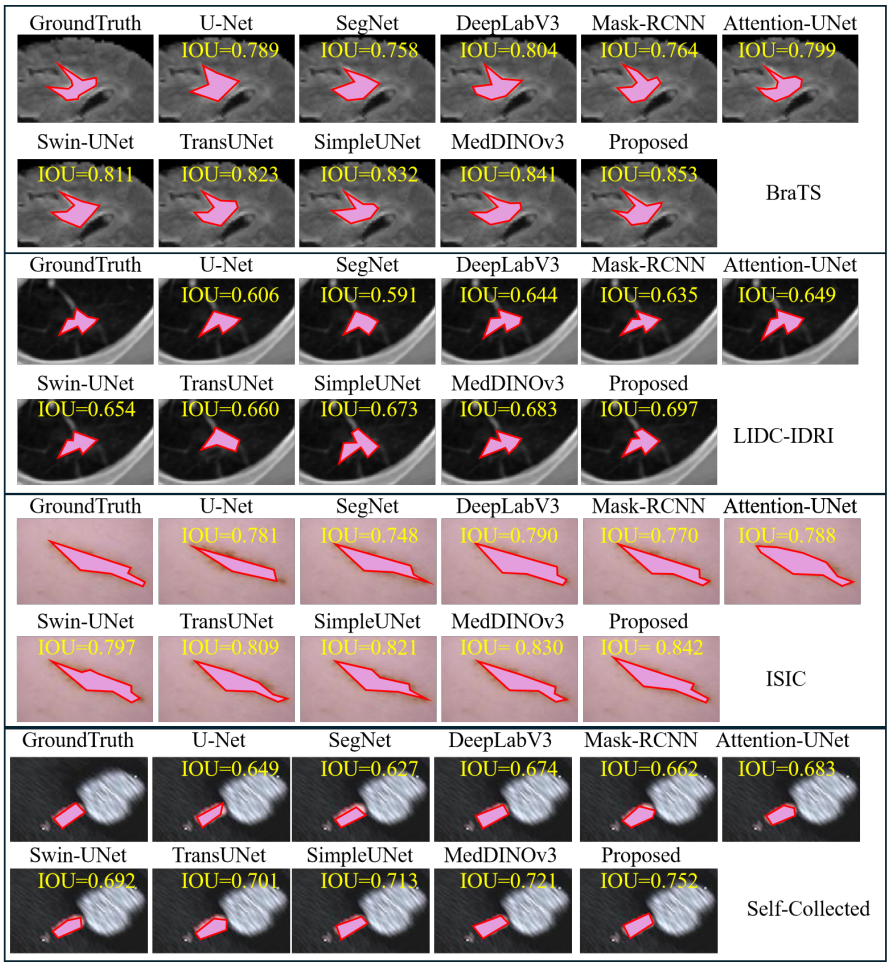


Figure 8. Visualization of different models' performance across different datasets.

4.3. Ablation Study

The ablation experiments were conducted to comprehensively evaluate the contributions of the local detail attention (LDA), global context attention (GCA), and cross-scale consistency (CS) modules in enhancing segmentation performance across different imaging modalities. Using the U-Net backbone as the baseline, each component was progressively added, and the performance variations were analyzed on the BraTS (MRI), LIDC-IDRI (CT), ISIC (dermoscopy), and self-collected datasets.

As shown in Table 5, the baseline model achieved reasonable segmentation accuracy but demonstrated noticeable limitations in Dice and IoU, especially in handling small-scale lesions and blurred or irregular boundaries. The introduction of LDA notably improved the model's ability to capture fine-grained boundary and texture information, leading to consistent gains in Precision and Recall across all datasets. This indicates that local attention effectively strengthens the model's sensitivity to subtle lesion structures. When GCA was further incorporated, the performance continued to improve, particularly in Dice and IoU, highlighting the advantage of integrating global contextual information. The GCA mechanism mitigates the locality constraints of convolutional features by modeling long-range dependencies, thereby enhancing structural coherence and segmentation stability in large or complex lesions. Finally, adding CS led to the best overall performance on all four datasets. This module enforces prediction consistency across multiple scales, acting as a regularization term that stabilizes optimization and reduces boundary-level prediction drift. The combination of all three components yielded consistent improvements across metrics and datasets, confirming that local refinement, global contextual modeling, and cross-scale regularization function synergistically to enhance both accuracy and robustness.

Table 5. An ablation study on the contributions of local detail attention (LDA), global context attention (GCA), and cross-scale consistency (CS) across different datasets. Metrics include Dice, IoU, Precision, Recall, and F1.

Dataset	Configuration	Dice	IoU	Precision	Recall	F1
BraTS	Baseline (U-Net backbone)	0.883	0.789	0.895	0.872	0.883
	+ LDA	0.897	0.806	0.907	0.886	0.897
	+ LDA + GCA	0.910	0.829	0.918	0.902	0.910
	+ LDA + GCA + CS (Proposed)	**0.923**	**0.852**	**0.931**	**0.916**	**0.923**
LIDC-IDRI	Baseline (U-Net backbone)	0.751	0.606	0.774	0.731	0.751
	+ LDA	0.770	0.632	0.796	0.751	0.770
	+ LDA + GCA	0.800	0.669	0.820	0.781	0.800
	+ LDA + GCA + CS (Proposed)	**0.823**	**0.696**	**0.841**	**0.809**	**0.823**
ISIC	Baseline (U-Net backbone)	0.877	0.782	0.893	0.863	0.877
	+ LDA	0.889	0.799	0.904	0.876	0.889
	+ LDA + GCA	0.903	0.819	0.915	0.890	0.903
	+ LDA + GCA + CS (Proposed)	**0.915**	**0.841**	**0.923**	**0.907**	**0.915**
Self-Collected	Baseline (U-Net backbone)	0.813	0.675	0.835	0.791	0.813
	+ LDA	0.831	0.700	0.850	0.808	0.831
	+ LDA + GCA	0.857	0.733	0.874	0.834	0.857
	+ LDA + GCA + CS (Proposed)	**0.871**	**0.753**	**0.888**	**0.849**	**0.871**

Note: Bold values indicate the best performance within each dataset.

From a theoretical perspective, LDA introduces adaptive a spatial–channel reweighting approach that selectively amplifies discriminative boundary features while suppressing background noise. GCA provides a global structural prior through low-rank attention modeling, ensuring semantic coherence across distant regions. CS, in turn, constrains predictions at different resolutions to converge toward a stable representation. Together, these

mechanisms contribute to a unified framework that enhances representational capacity and generalization ability in multi-modal medical image segmentation tasks.

4.4. Discussion

In practical medical imaging applications, the hierarchical attention framework demonstrates significant clinical value and application potential. For instance, in brain tumor MRI, clinicians must not only estimate tumor volume and boundary but also differentiate among enhancing regions, necrotic areas, and edema. Traditional approaches often produce misjudgments under conditions of blurred boundaries or complex lesions, whereas the proposed model establishes a synergistic relationship between global context and local details, producing segmentation results that align more closely with clinical requirements. This capability supports faster and more accurate surgical planning and therapeutic evaluation. In lung nodule detection with CT images, the framework is particularly effective for identifying nodules smaller than 10 mm, emphasizing texture details under low-contrast conditions and reducing missed diagnoses, thereby providing a reliable computer-aided diagnostic tool for early lung cancer screening. In dermatology, physicians frequently rely on dermoscopic imaging to evaluate pigmented lesions and determine their malignancy potential. By enhancing boundary and chromatic differentiation features, the proposed method achieves high-precision segmentation and classification, even under complex backgrounds and variable lighting, which contributes to the early detection of melanoma. In real-world medical workflows, the model's application extends beyond academic segmentation tasks. Within radiology workstations, it can be embedded into daily reading systems to automatically annotate suspicious lesion regions, reducing the burden of frame-by-frame inspection in large image datasets. In longitudinal monitoring, the model facilitates comparison across multiple time points, enabling clinicians to evaluate lesion progression or therapeutic effects more intuitively, thus supporting personalized treatment planning. In telemedicine and resource-limited regions, this model can be deployed on imaging platforms in smaller hospitals, compensating for the shortage of specialized radiologists and enhancing diagnostic efficiency and capacity in primary healthcare institutions. From a medical economics perspective, the deployment of such an automated and scalable segmentation framework can substantially reduce diagnostic costs by minimizing the time and labor required for image interpretation. By improving workflow efficiency and reducing the dependency on highly specialized experts, especially in resource-limited hospitals, the model contributes to more equitable access to diagnostic services and promotes cost-effective healthcare delivery. These advantages highlight the potential of intelligent imaging systems not only to enhance diagnostic accuracy but also to improve the overall economic sustainability of healthcare systems.

4.5. Limitation and Future Work

Although the proposed hierarchical attention mechanism achieves outstanding performance in multimodal medical image segmentation and recognition, several limitations remain to be addressed. The experimental data were primarily derived from public datasets and partially constructed image collections. Despite covering multiple modalities such as MRI, CT, and dermoscopy, the sample size and disease distribution still diverge from real clinical settings, especially in the representation of rare lesions and multi-center consistency, which may affect the generalizability of the model. Moreover, the introduction of multi-level attention and cross-scale consistency increases computational and storage overhead, posing challenges for deployment in resource-constrained medical institutions or portable devices. In addition, the interpretability of the model remains limited. Although attention mechanisms provide certain visualization cues, the internal decision-making process does

not yet fully align with clinicians' diagnostic reasoning, potentially affecting trust and widespread adoption. Future research will focus on two key directions. First, larger, multi-center, and multimodal clinical datasets should be incorporated to enhance robustness and adaptability across regions and devices. Second, model compression and lightweight design will be explored for efficient deployment in edge devices and real-time applications, alongside the development of improved interpretability mechanisms that can better align model outputs with clinical judgment. Through these efforts, the framework is expected to be further advanced toward integration into real-world clinical workflows and practical deployment.

5. Conclusions

The present study focuses on critical challenges in medical image segmentation and recognition, particularly addressing issues such as blurred lesion boundaries, large scale variations, and the difficulty of modeling long-range dependencies across different modalities. A unified framework based on hierarchical attention mechanisms is proposed to mitigate these challenges. Experimental results demonstrate that, on the integrated dataset, the proposed method outperforms both traditional convolutional models and state-of-the-art Transformer-based approaches across multiple metrics including Dice, IoU, Precision, Recall, and F1. The overall accuracy reaches Dice = 0.886, Precision = 0.898, and Recall = 0.875, highlighting its comprehensive superiority. Ablation studies further validate the contribution of each submodule: local detail modeling strengthens boundary and texture features, global context modeling compensates for the limited receptive field of convolutional operations, and cross-scale consistency ensures stability and coherence across predictions at different resolutions. In summary, this study not only introduces an innovative hierarchical attention mechanism but also demonstrates its outstanding performance and broad applicability through extensive experiments, thereby providing new solutions and technical directions for intelligent analysis of medical images. Beyond its technical contributions, the proposed framework also shows potential real-world benefits in medical economics by streamlining diagnostic workflows, reducing redundant imaging costs, and supporting the more efficient allocation of healthcare resources.

Author Contributions: Conceptualization, Y.Z. (Yi Zhu), Y.Z. (Yawen Zhu), H.M. and M.L.; Data curation, B.L. and X.W.; Formal analysis, L.X.; Funding acquisition, M.L.; Investigation, L.X.; Methodology, Y.Z. (Yi Zhu), Y.Z. (Yawen Zhu) and H.M.; Project administration, M.L.; Resources, B.L. and X.W.; Software, Y.Z. (Yi Zhu), Y.Z. (Yawen Zhu) and H.M.; Supervision, M.L.; Validation, L.X.; Visualization, B.L. and X.W.; Writing—original draft, Y.Z. (Yi Zhu), Y.Z. (Yawen Zhu), H.M., B.L., L.X., X.W. and M.L.; Y.Z. (Yi Zhu), Y.Z. (Yawen Zhu), and H.M. contributed equally to this work. All authors have read and agreed to the published version of this manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61202479.

Data Availability Statement: The original data presented in this study are available on request from the corresponding author due to privacy and institutional restrictions. The code and related materials are openly available in GitHub at https://github.com/Aurelius-04/Med_Seg.git (accessed on 28 October 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, Y.; Shen, Z.; Jiao, R. Segment anything model for medical image segmentation: Current applications and future directions. *Comput. Biol. Med.* **2024**, *171*, 108238. [CrossRef]
2. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical applications of artificial intelligence in medical imaging and image processing—A review. *Cancers* **2024**, *16*, 1870. [CrossRef]
3. Berghout, T. The neural frontier of future medical imaging: A review of deep learning for brain tumor detection. *J. Imaging* **2024**, *11*, 2. [CrossRef]
4. Aparnaa, R.; Dinesh, R. Multi-disease diagnosis using medical images. In Proceedings of the 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), Namakkal, India, 15–16 March 2024; pp. 1–6.
5. Li, X.; Zhang, L.; Yang, J.; Teng, F. Role of artificial intelligence in medical image analysis: A review of current trends and future directions. *J. Med Biol. Eng.* **2024**, *44*, 231–243. [CrossRef]
6. Li, H.; Bu, Q.; Shi, X.; Xu, X.; Li, J. Non-invasive medical imaging technology for the diagnosis of burn depth. *Int. Wound J.* **2024**, *21*, e14681. [CrossRef]
7. Baştuğ, B.T.; Başol, H.G. Radiology's role in dermatology: A closer look at two years of data. *Eur. Res. J.* **2025**, *11*, 395–403. [CrossRef]
8. Vavekanand, R. A deep learning approach for medical image segmentation integrating magnetic resonance imaging to enhance brain tumor recognition. *SSRN* **2024**, 4827019, *in press*. [CrossRef]
9. Xu, Y.; Quan, R.; Xu, W.; Huang, Y.; Chen, X.; Liu, F. Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering* **2024**, *11*, 1034. [CrossRef] [PubMed]
10. Mostafa, R.R.; Houssein, E.H.; Hussien, A.G.; Singh, B.; Emam, M.M. An enhanced chameleon swarm algorithm for global optimization and multi-level thresholding medical image segmentation. *Neural Comput. Appl.* **2024**, *36*, 8775–8823. [CrossRef]
11. Zhou, X.; Chen, S.; Ren, Y.; Zhang, Y.; Fu, J.; Fan, D.; Lin, J.; Wang, Q. Atrous Pyramid GAN Segmentation Network for Fish Images with High Performance. *Electronics* **2022**, *11*, 911. [CrossRef]
12. Azad, R.; Aghdam, E.K.; Rauland, A.; Jia, Y.; Avval, A.H.; Bozorgpour, A.; Karimijafarbigloo, S.; Cohen, J.P.; Adeli, E.; Merhof, D. Medical image segmentation review: The success of u-net. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10076–10095. [CrossRef] [PubMed]
13. Liu, L.; Li, X.; Wang, S.; Wang, J.; Melo, S.N. MRCA-UNet: A Multiscale Recombined Channel Attention U-Net Model for Medical Image Segmentation. *Symmetry* **2025**, *17*, 892. [CrossRef]
14. Huang, L.; Miron, A.; Hone, K.; Li, Y. Segmenting medical images: From UNet to res-UNet and nnUNet. In Proceedings of the 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), Guadalajara, Mexico, 26–28 June 2024; pp. 483–489.
15. Chen, B.; Li, Y.; Liu, J.; Yang, F.; Zhang, L. MSMHSA-DeepLab V3+: An Effective Multi-Scale, Multi-Head Self-Attention Network for Dual-Modality Cardiac Medical Image Segmentation. *J. Imaging* **2024**, *10*, 135. [CrossRef] [PubMed]
16. Xiong, L.; Yi, C.; Xiong, Q.; Jiang, S. SEA-NET: Medical image segmentation network based on spiral squeeze-and-excitation and attention modules. *BMC Med. Imaging* **2024**, *24*, 17. [CrossRef]
17. Zheng, F.; Chen, X.; Liu, W.; Li, H.; Lei, Y.; He, J.; Pun, C.M.; Zhou, S. Smaformer: Synergistic multi-attention transformer for medical image segmentation. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 4048–4053.
18. Zhang, C.; Sun, S.; Hu, W.; Zhao, P. FDR-TransUNet: A novel encoder-decoder architecture with vision transformer for improved medical image segmentation. *Comput. Biol. Med.* **2024**, *169*, 107858. [CrossRef]
19. Chen, J.; Zhang, X.; Li, R.; Zhou, P. Swin-HAUnet: A Swin-Hierarchical Attention Unet For Enhanced Medical Image Segmentation. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Urumqi, China, 18–20 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 371–385.
20. Ghamsarian, N.; Wolf, S.; Zinkernagel, M.; Schoeffmann, K.; Sznitman, R. Deeppyramid+: Medical image segmentation using pyramid view fusion and deformable pyramid reception. *Int. J. Comput. Assist. Radiol. Surg.* **2024**, *19*, 851–859. [CrossRef]
21. Hikmah, N.F.; Hajjanto, A.D.; Surbakti, A.F.A.; Prakosa, N.A.; Asmaria, T.; Sardjono, T.A. Brain tumor detection using a MobileNetV2-SSD model with modified feature pyramid network levels. *Int. J. Electr. Comput. Eng.* **2024**, *14*, 3995–4004. [CrossRef]
22. Han, X.; Li, T.; Bai, C.; Yang, H. Integrating prior knowledge into a bibranch pyramid network for medical image segmentation. *Image Vis. Comput.* **2024**, *143*, 104945. [CrossRef]
23. Chen, A.; Wei, Y.; Le, H.; Zhang, Y. Learning by teaching with ChatGPT: The effect of teachable ChatGPT agent on programming education. *Br. J. Educ. Technol.* 2024, *in press*. [CrossRef]
24. Kumar, R.R.; Priyadarshi, R. Denoising and segmentation in medical image analysis: A comprehensive review on machine learning and deep learning approaches. *Multimed. Tools Appl.* **2025**, *84*, 10817–10875. [CrossRef]

25. Zhang, Y.; Mao, Y.; Lu, X.; Zou, X.; Huang, H.; Li, X.; Li, J.; Zhang, H. From single to universal: Tiny lesion detection in medical imaging. *Artif. Intell. Rev.* **2024**, *57*, 192. [CrossRef]
26. Krithika Alias AnbuDevi, M.; Suganthi, K. Review of semantic segmentation of medical images using modified architectures of UNET. *Diagnostics* **2022**, *12*, 3064. [CrossRef]
27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
28. Maji, D.; Sigedar, P.; Singh, M. Attention Res-UNet with Guided Decoder for semantic segmentation of brain tumors. *Biomed. Signal Process. Control* **2022**, *71*, 103077. [CrossRef]
29. Ali, S.; Khurram, R.; Rehman, K.u.; Yasin, A.; Shaukat, Z.; Sakhawat, Z.; Mujtaba, G. An improved 3D U-Net-based deep learning system for brain tumor segmentation using multi-modal MRI. *Multimed. Tools Appl.* **2024**, *83*, 85027–85046. [CrossRef]
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
31. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
32. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef] [PubMed]
33. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 205–218.
34. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
35. Gao, Y.; Zhou, M.; Metaxas, D.N. UTNet: A hybrid transformer architecture for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 61–71.
36. Tian, Q.; Wang, Z.; Cui, X. Improved Unet brain tumor image segmentation based on GSConv module and ECA attention mechanism. *arXiv* **2024**, arXiv:2409.13626. [CrossRef]
37. Zhao, P.; Li, Z.; You, Z.; Chen, Z.; Huang, T.; Guo, K.; Li, D. SE-U-lite: Milling tool wear segmentation based on lightweight U-net model with squeeze-and-excitation module. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–8. [CrossRef]
38. Matlala, B.; van der Haar, D.; Vadapalli, H. Automated Gross Tumor Volume Segmentation in Meningioma Using Squeeze and Excitation Residual U-Net for Enhanced Radiotherapy Planning. In Proceedings of the International Conference on AI in Healthcare, Basel, Switzerland, 10–11 November 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 57–67.
39. Wu, Y.; Lin, Y.; Xu, T.; Meng, X.; Liu, H.; Kang, T. Multi-Scale Feature Integration and Spatial Attention for Accurate Lesion Segmentation. In Proceedings of the 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, 20–22 June 2025.
40. Chen, Y.; Zou, B.; Guo, Z.; Huang, Y.; Huang, Y.; Qin, F.; Li, Q.; Wang, C. Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 7759–7767.
41. Chen, G.; Zhou, L.; Zhang, J.; Yin, X.; Cui, L.; Dai, Y. ESKNet: An enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation. *Expert Syst. Appl.* **2024**, *246*, 123265. [CrossRef]
42. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
43. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
44. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
45. Liu, J.; Zhao, D.; Shen, J.; Geng, P.; Zhang, Y.; Yang, J.; Zhang, Z. HRD-Net: High resolution segmentation network with adaptive learning ability of retinal vessel features. *Comput. Biol. Med.* **2024**, *173*, 108295. [CrossRef]
46. Dai, H.; Xie, W.; Xia, E. SK-Unet++: An improved Unet++ network with adaptive receptive fields for automatic segmentation of ultrasound thyroid nodule images. *Med. Phys.* **2024**, *51*, 1798–1811. [CrossRef]
47. Nguyen, H.T.; Nguyen, N.M.; Huynh, T.Q.; Su, A.K. An enhanced UNet3+ model for accurate identification of COVID-19 in CT images. *Digit. Signal Process.* **2025**, *163*, 105205. [CrossRef]

48. De Verdier, M.C.; Saluja, R.; Gagnon, L.; LaBella, D.; Baid, U.; Tahon, N.H.; Foltyn-Dumitru, M.; Zhang, J.; Alafif, M.; Baig, S.; et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv* **2024**, arXiv:2405.18368.
49. Armato III, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef] [PubMed]
50. Codella, N.C.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172.
51. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med Imaging* **2014**, *34*, 1993–2024. [CrossRef]
52. Kalpathy-Cramer, J.; Zhao, B.; Goldgof, D.; Gu, Y.; Wang, X.; Yang, H.; Tan, Y.; Gillies, R.; Napel, S. A comparison of lung nodule segmentation algorithms: Methods and results from a multi-institutional study. *J. Digit. Imaging* **2016**, *29*, 476–487. [CrossRef]
53. Yeap, P.L.; Wong, Y.M.; Ong, A.L.K.; Tuan, J.K.L.; Pang, E.P.P.; Park, S.Y.; Lee, J.C.L.; Tan, H.Q. Predicting dice similarity coefficient of deformably registered contours using Siamese neural network. *Phys. Med. Biol.* **2023**, *68*, 155016. [CrossRef]
54. He, Y.; Zhang, N.; Ge, X.; Li, S.; Yang, L.; Kong, M.; Guo, Y.; Lv, C. Passion Fruit Disease Detection Using Sparse Parallel Attention Mechanism and Optical Sensing. *Agriculture* **2025**, *15*, 733. [CrossRef]
55. Zhou, C.; Ge, X.; Chang, Y.; Wang, M.; Shi, Z.; Ji, M.; Wu, T.; Lv, C. A Multimodal Parallel Transformer Framework for Apple Disease Detection and Severity Classification with Lightweight Optimization. *Agronomy* **2025**, *15*, 1246. [CrossRef]
56. Li, L.; Ma, H.; Zhang, X.; Zhao, X.; Lv, M.; Jia, Z. Synthetic aperture radar image change detection based on principal component analysis and two-level clustering. *Remote Sens.* **2024**, *16*, 1861. [CrossRef]
57. Li, L.; Ma, H.; Jia, Z. Multiscale geometric analysis fusion-based unsupervised change detection in remote sensing images via FLICM model. *Entropy* **2022**, *24*, 291. [CrossRef]
58. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
59. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
60. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [CrossRef]
61. Yu, X.; Chen, Y.; He, G.; Zeng, Q.; Qin, Y.; Liang, M.; Luo, D.; Liao, Y.; Ren, Z.; Kang, C.; et al. Simple is what you need for efficient and accurate medical image segmentation. *arXiv* **2025**, arXiv:2506.13415. [CrossRef]
62. Li, Y.; Wu, Y.; Lai, Y.; Hu, M.; Yang, X. MedDINOv3: How to adapt vision foundation models for medical image segmentation? *arXiv* **2025**, arXiv:2509.02379. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.