

Article

Hybrid CNN–Transformer with Fusion Discriminator for Ovarian Tumor Ultrasound Imaging Classification

Donglei Xu ^{1,2}, Xinyi He ², Ruoyun Zhang ², Yinuo Zhang ², Manzhou Li ² and Yan Zhan ^{2,3,*}¹ Department of Gynecology, The Peoples Hospital of Langfang City, Langfang 065000, China² National School of Development, Peking University, Beijing 100871, China³ Artificial Intelligence Research Institute, Tsinghua University, Beijing 100084, China

* Correspondence: yan_zhan_thu@163.com

Abstract

We propose a local–global attention fusion network for benign–malignant discrimination of ovarian tumors in color Doppler ultrasound (CDFI). The framework integrates three complementary modules: a local enhancement module (LEM) to capture fine-grained texture and boundary cues, a Global Attention Module (GAM) to model long-range dependencies with flow-aware priors, and a Fusion Discriminator (FD) to align and adaptively reweight heterogeneous evidence for robust decision-making. The method was evaluated on a multi-center clinical dataset comprising 820 patient cases (482 benign and 338 malignant), ensuring a realistic and moderately imbalanced distribution. Compared with classical baselines including ResNet-50, DenseNet-121, ViT, Hybrid CNN–Transformer, U-Net, and SegNet, our approach achieved an accuracy of 0.923, sensitivity of 0.911, specificity of 0.934, AUC of 0.962, and F1-score of 0.918, yielding improvements of about three percentage points in the AUC and F1-score over the strongest baseline. Ablation experiments confirmed the necessity of each module, with the performance degrading notably when the GAM or the LEM was removed, while the complete design provided the best results, highlighting the benefit of local–global synergy. Five-fold cross-validation further demonstrated stable generalization (accuracy: 0.922; AUC: 0.961). These findings indicate that the proposed system offers accurate and robust assistance for preoperative triage, surgical decision support, and follow-up management of ovarian tumors.



Academic Editor: Luca Mesin

Received: 15 September 2025

Revised: 1 October 2025

Accepted: 9 October 2025

Published: 14 October 2025

Citation: Xu, D.; He, X.; Zhang, R.; Zhang, Y.; Li, M.; Zhan, Y. Hybrid CNN–Transformer with Fusion Discriminator for Ovarian Tumor Ultrasound Imaging Classification. *Electronics* **2025**, *14*, 4040. <https://doi.org/10.3390/electronics14204040>

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/>).

Keywords: machine learning; ovarian tumor imaging classification; Hybrid CNN–Transformer; ultrasound imaging

1. Introduction

Ovarian tumor malignancy assessment plays a crucial role in gynecological clinical diagnosis and therapeutic decision-making. Color Doppler flow imaging (CDFI), owing to its non-invasive, real-time, and low-cost properties, has become the preferred imaging modality in clinical practice. By detecting blood flow signals, it provides auxiliary diagnostic information; however, CDFI images suffer from blurred boundaries, noise interference, and complex hemodynamic patterns, leading to strong dependence on physician expertise and unstable diagnostic accuracy. For example, abundant blood flow signals may suggest malignancy, yet similar manifestations can also occur in benign tumors, thereby increasing diagnostic difficulty [1].

In recent years, deep learning has brought significant breakthroughs to medical image analysis. Convolutional neural networks (CNNs), through multilayer convolution operations, can automatically extract local features such as textures and edges [2], achieving

superior performance compared with conventional methods in ovarian tumor classification. For instance, models such as ResNet-50 have achieved a classification accuracy exceeding 85% on public datasets [3]. Nevertheless, most existing approaches primarily focus on single-feature extraction, such as CNN-based texture analysis, and fail to effectively integrate both local structural details and global blood flow distribution into CDFI images [4]. This limitation becomes particularly evident when handling images with complex backgrounds and blurred lesion boundaries, impairing diagnostic accuracy and robustness [5]. Specifically, U-Net and its variants exhibit a restricted capability to capture global features, which diminishes the segmentation accuracy, especially for images with high foreground–background similarity. Moreover, while the translation invariance of CNNs enhances model generalization, it also reduces the sensitivity to positional information, thereby limiting the segmentation performance for irregularly shaped lesions. Regarding hemodynamic modeling, most segmentation algorithms mainly emphasize anatomical information while underexploring blood flow signals. In low-contrast ultrasound images, the current models still struggle with lesion delineation at blurred boundaries, restricting the accurate modeling of complex blood flow patterns [6]. Additionally, the local receptive field of CNNs makes it challenging to capture long-range dependencies in blood flow signals, and the recognition of small lesions with indistinct boundaries (e.g., solid nodules smaller than 1 cm) remains limited [3,7]. Vision Transformers (ViTs), by leveraging self-attention to model the global context, significantly improve the recognition accuracy under complex backgrounds in breast ultrasound nodule classification, yet their sensitivity to local details (e.g., microcalcifications within lesions) remains insufficient, and the computational complexity is high [8].

Ultrasound combined with CDFI achieves relatively high accuracy in distinguishing benign from malignant ovarian tumors, aiding early diagnosis and treatment, yet CDFI alone is insufficient and must be complemented with grayscale morphology. Although morphological indices with Doppler parameters showed good sensitivity and specificity, Doppler flow added little accuracy, and indices such as the resistance index had limitations [9]; moreover, overlaps between complex benign and malignant lesions [10] and the presence of blood flow in both types of tumors [6] reduce reliability. These diagnostic shortcomings risk misdiagnosis or repeated examinations, leading to disease progression, unnecessary invasive treatments, and increased healthcare burden [11]. To overcome these challenges, we propose a local–global attention fusion model for lesion detection and classification in CDFI images, combining CNNs for local feature extraction with Transformers for global context modeling to enhance malignancy prediction. This approach shortens diagnostic pathways, reduces redundant exams, and lessens the reliance on costly MRI/CT, thereby alleviating economic pressure on patients and healthcare systems. The main contributions are summarized as follows:

1. A novel deep network architecture based on local–global attention fusion is introduced to enhance tumor boundary recognition in CDFI images.
2. A local enhancement module is designed to highlight fine-grained lesion features.
3. A global attention mechanism is incorporated to capture long-range dependencies between blood flow signals and lesion regions.
4. The superiority of the proposed model is validated on multi-center CDFI datasets, achieving substantial improvements over classical networks such as ResNet and DenseNet.
5. An economic analysis framework is proposed to investigate how the model alleviates insufficient healthcare resources (e.g., shortage of experienced sonographers). Through cost-effectiveness simulation, scalability analysis, and social value evalua-

tion, the framework provides a comprehensive perspective on the economic impact of early diagnosis for both patients and society.

2. Related Work

2.1. Ultrasound Imaging Analysis of Ovarian Tumor Malignancy

In ovarian tumor detection and malignancy prediction based on CDFI, traditional feature-based methods relying on morphological and hemodynamic parameters suffer from strong subjectivity and limited accuracy [12]; thresholding, region-growing, and edge-based methods are highly noise-sensitive and prone to over-segmentation, impairing lesion delineation [13]. Deep learning has improved performance, with CNNs and U-Net variants achieving Dice scores above 0.83 through spatial and channel attention [14], yet challenges remain in modeling long-range dependencies and segmenting blurred or low-contrast lesions [15,16]. Moreover, while ultrasound combined with CDFI shows relatively high accuracy for distinguishing benign from malignant tumors, CDFI alone is insufficient: Doppler parameters add little diagnostic gain, resistance index values are limited [17], and overlaps between benign and malignant features—as well as the blood flow present in both—reduce reliability [9].

2.2. Applications of Deep Learning in Medical Ultrasound

CNNs and U-Net-based networks are widely applied in ultrasound image segmentation and classification [18,19], but their fixed receptive fields limit global feature modeling [20], leading to reduced accuracy in complex backgrounds, blurred boundaries, and irregular lesion shapes [21]. Most existing methods also underexplore hemodynamic signals, showing deficiencies in modeling blood flow complexity under low-contrast conditions [22], while CDFI alone remains diagnostically limited. Recently, Transformer architectures have been introduced to capture the global context via self-attention [23], and multi-head self-attention has improved the analysis of correlations between blood flow distribution and lesion activity [24]. These advances suggest that hybrid CNN–Transformer models can integrate local details with global semantics, enhancing the analysis of complex medical images.

2.3. Attention Mechanisms and Global Modeling Methods

The application of Transformers and attention mechanisms has provided new strategies for medical image analysis, where self-attention effectively captures global context [25] and multi-head self-attention models hemodynamic–metabolic relationships [26], improving the recognition of malignant blood flow patterns. Frameworks such as TransMed, the first Transformer-based multimodal classification model, integrated self-attention for multimodal fusion [27], while MTS-Net incorporated dual-enhanced positional multi-head self-attention to boost 3D CT diagnosis of May–Thurner syndrome [28]. These studies confirmed the advantages of Transformer–CNN hybrids in utilizing global and local information [29], laying the foundation for local–global attention fusion models [30]. However, challenges remain, including high computational cost ($O(n^2)$ for a sequence length n) [31], the reliance on large annotated datasets, and difficulties handling multimodal distributional differences [32,33]. Common solutions such as downsampling or patching risk losing critical details [34], highlighting the need for efficient architectures and robust fusion strategies in medical imaging.

2.4. Medical Artificial Intelligence and Health Economics

The application of artificial intelligence in healthcare (AIH) has gradually transcended the pursuit of algorithmic accuracy, becoming a central topic in health economics re-

search [35]. Prior studies have demonstrated that intelligent imaging systems offer unique advantages in cost containment, efficiency improvement, and equitable resource allocation [36]. On the one hand, intelligent diagnostic tools can substantially reduce costs by decreasing physician reading times, minimizing repeated examinations, and shortening diagnostic pathways [37]. In early cancer screening and diagnostic assistance scenarios, AI systems have been shown to lower per-unit diagnostic costs and, in some studies, achieve a cost-effectiveness comparable to or surpassing that of experienced radiologists [38]. On the other hand, AI systems can provide consistent diagnostic capabilities in primary care and resource-limited regions, partially compensating for the uneven distribution of medical expertise and advanced imaging equipment, thereby improving fairness in healthcare resource allocation on a larger scale [39]. From an economic perspective, the value of medical AI depends not only on diagnostic accuracy but also on economic feasibility and clinical applicability. If an AI system maintains high accuracy while operating at low costs, it can be widely adopted for high-frequency applications such as routine screening, chronic disease management, and long-term monitoring. Conversely, systems requiring high computational resources and complex deployment, despite their excellent performance under experimental conditions, may not be scalable in clinical practice [40]. Consequently, increasing emphasis has been placed on incorporating cost-effectiveness analysis (CEA) and health economic modeling into AI evaluation [41], providing comprehensive assessments of its value for patients, healthcare institutions, and insurance systems [42].

3. Materials and Method

3.1. Data Collection

In this study, clinical and imaging data were obtained from open-source datasets on patients with ovarian tumors who underwent surgical treatment at Langfang People's Hospital, Hebei Province, China, between 2018 and 2024. All cases were pathologically confirmed postoperatively and included both benign and malignant ovarian tumors. According to surgical procedures, patients with benign ovarian tumors primarily underwent umbilical + 1-port laparoscopic surgery, while patients with malignant ovarian tumors were predominantly treated with laparotomy, as shown in Table 1 and Figure 1. Preoperative imaging data were collected by experienced sonographers and included grayscale B-mode ultrasound images and CDFI scans. To ensure quality and consistency, all examinations were performed by senior sonographers with more than 10 years of clinical experience using identical ultrasound diagnostic equipment. During acquisition, patients were positioned supine or with a filled bladder to optimize visualization of the pelvic structures. Grayscale B-mode images recorded tumor size, morphology, capsular characteristics, and internal echogenicity, whereas CDFI images captured the intra- and peritumoral blood flow distribution and vascularity. All images were stored in DICOM format at a standardized resolution of 512×512 pixels and were anonymized to remove patient identifiers.

Table 1. Data distribution of ovarian tumor cases collected from Langfang People's Hospital (2018–2024).

Category	Surgical Approach	Number of Cases
Benign ovarian tumors	Umbilical + 1-port laparoscopy	426
Malignant ovarian tumors	Laparotomy	394

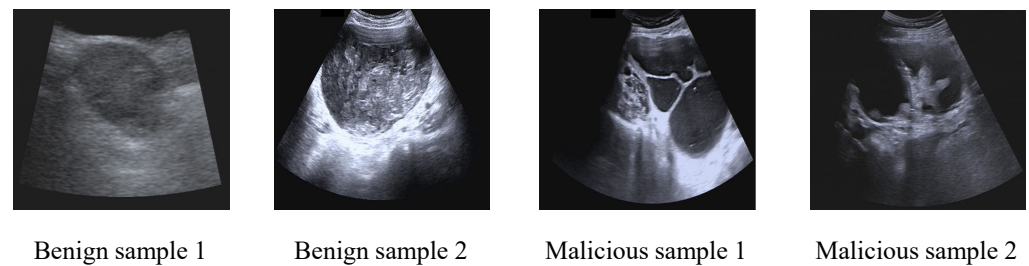


Figure 1. Dataset overview. Representative CDFI and grayscale ultrasound images of ovarian masses: Benign Sample 1 and Malignant Sample 1 show CDFI images from different patients, while Benign Sample 2 and Malignant Sample 2 show grayscale ultrasound images from different patients.

During data acquisition, for benign ovarian tumors, emphasis was placed on features such as cystic component proportion, capsular smoothness, the presence of septations, and blood flow intensity. For malignant ovarian tumors, greater attention was given to the solid component proportion, papillary projections, intratumoral necrosis, and irregular hemodynamic patterns. Each patient underwent ultrasound examination within one week prior to surgery to ensure temporal consistency between the imaging findings and pathological outcomes. Postoperative pathological diagnoses were independently reviewed and confirmed by two senior pathologists, serving as the final reference standard. In total, a defined number of benign ovarian tumors and ovarian cancer cases was included. A retrospective analysis of their preoperative ultrasound features was performed to summarize imaging differences between benign and malignant tumors, thereby providing a reliable foundation for the proposed local–global attention fusion model. The collected dataset ensured both the diversity and completeness of the imaging data, encompassing the entire clinical pathway from preoperative examination to postoperative pathology, and established a solid basis for model training and validation.

3.2. Data Preprocessing and Augmentation

Data preprocessing represents a critical step in medical image analysis and exerts a substantial influence on model performance. In this study, comprehensive preprocessing and augmentation were applied to CDFI images to ensure that the model could learn informative features from high-quality data. Furthermore, augmentation strategies including random rotation, flipping, CutMix, and brightness adjustment were employed to enhance model generalization, thereby ensuring stability and robustness under varying data distributions. These techniques effectively expanded the diversity of the training dataset and improved the adaptability to different imaging scenarios. The principles of these preprocessing and augmentation methods, along with their mathematical formulations, are described as follows.

Noise reduction was first performed. As CDFI images are often affected by noise, a combined approach using Gaussian and median filtering was adopted. Gaussian filtering smoothed the image through convolution operations, reducing high-frequency noise, while median filtering replaced each pixel with the median value of its neighborhood, effectively eliminating salt-and-pepper noise. The combination of these methods preserved important image details while reducing noise interference. Subsequently, pseudo-color channel separation and normalization were conducted. CDFI images contain two primary channels, namely grayscale B-mode and blood flow signals. These channels were separated and individually normalized. For each channel, zero-mean and unit-variance normalization was applied, making the data suitable for neural network input. The normalization was defined as

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

where I denotes the raw pixel value, μ and σ represent the mean and standard deviation of the image, and I_{norm} is the normalized output.

Following image preprocessing, textual data standardization was performed for multimodal graph data containing textual information. In graph neural network applications, node and edge attributes often include textual descriptors in addition to numerical features. Standardization of textual data typically involves stop-word removal, lemmatization, tokenization, and vectorization. Initially, non-essential words were eliminated to reduce noise. Then, lemmatization or stemming ensured consistent representation of semantically similar terms. Finally, the text was transformed into high-dimensional vectors using approaches such as the bag-of-words model or Word2Vec. The vectorization process was formulated as

$$\mathbf{v}_t = \text{Word2Vec}(t), \quad (2)$$

where $\mathbf{v}_t$ denotes the vector representation of token t , and $\text{Word2Vec}(t)$ refers to the mapping operation. Graph structure cleaning was subsequently performed to remove redundant or irrelevant nodes and edges, thereby improving representation efficiency and simplifying subsequent computations. Common procedures included eliminating isolated nodes and redundant edges. The cleaned graph was defined as

$$G_{\text{clean}} = \{V, E \mid |V| \geq 1, |E| \geq 1, \text{edges unique}\}, \quad (3)$$

where G_{clean} is the cleaned graph, V the set of nodes, and E the set of edges.

Edge-weight normalization was also critical in graph neural network training. Since edge weights represent relational strengths, normalization ensured balanced contributions during message passing. Min-max normalization was used to rescale edge weights to the interval $[0, 1]$, defined as

$$w_{\text{norm}} = \frac{w - \min(W)}{\max(W) - \min(W)}, \quad (4)$$

where w is the raw edge weight, W the set of all edge weights, and w_{norm} the normalized value. Graph augmentation techniques were further applied to increase data diversity and improve robustness. Edge sampling was implemented by randomly selecting a subset of edges to construct subgraphs:

$$G_{\text{sampled}} = \{V, E_s \mid E_s \subseteq E\}, \quad (5)$$

where G_{sampled} denotes the sampled subgraph and E_s the subset of edges. Perturbation introduced random variations into nodes and edges, expressed as

$$G_{\text{perturbed}} = \{V + \epsilon, E + \delta\}, \quad (6)$$

where ϵ and δ represent perturbations of nodes and edges, respectively. Local reconnection modified the graph topology by adding new edges:

$$E_{\text{reconnected}} = E \cup \{(v_i, v_j)\}, \quad (7)$$

where $E_{\text{reconnected}}$ is the updated edge set and (v_i, v_j) a newly added edge. In parallel, multiple augmentation techniques were applied to ultrasound images. Random rotation and flipping altered the orientation and position, simulating variable imaging angles and patient positions. CutMix combined patches from multiple images, generating compos-

ite samples to enhance the adaptability to diverse lesion morphologies and positions. Brightness adjustment simulated variations in imaging conditions by randomly modifying luminance, thereby improving robustness to contrast fluctuations.

3.3. Proposed Method

The overall methodological pipeline can be summarized as follows: standardized color Doppler ultrasound images are first input into the feature encoding stage of the model, where they are mapped into multi-dimensional tensor representations before entering different modules. The model architecture consists of three main components: the Global Attention Module (GAM), the local enhancement module (LEM), and the Fusion Discriminator (FD). These modules are sequentially connected to accomplish hierarchical feature extraction and aggregation. Specifically, a portion of the input features is directed into the LEM, where multilayer convolution operations extract local textures and boundary details of the lesions. Channel attention is then employed to emphasize responses from key regions, thereby forming fine-grained local feature representations. Another portion of the features is fed into the GAM, where the multi-head self-attention mechanism of the Transformer structure establishes long-range dependencies between pixels, capturing global interaction patterns between lesion regions and blood flow distribution. This process outputs feature representations enriched with contextual global semantics. During the fusion stage, the local features from the LEM and the global features from the GAM are projected into a unified representational space in the FD. Through concatenation or weighted fusion, complementary information is integrated. The fused high-dimensional features are then input into a multilayer perceptron (MLP), where nonlinear transformations and normalization further abstract them into more discriminative classification features. Finally, the output layer performs binary classification, predicting the benign or malignant label of the tumor. This end-to-end workflow achieves a direct mapping from the raw image input to the classification output, where the synergy of local and global information modeling ensures that the model attends to fine details while maintaining awareness of the overall blood flow patterns, thereby achieving robustness and high accuracy in complex imaging scenarios.

3.3.1. Local Enhancement Module

The core design of the LEM lies in thoroughly mining the fine-grained textures and boundary features of lesion regions in ultrasound images, thereby compensating for the limitations of global attention in capturing local details. As shown in Figure 2, this module is based on CNNs, employing stacked convolutional layers with nonlinear activations to achieve multi-scale feature extraction. Initially, a 3×3 convolution kernel is applied to mapping features from the input image, followed by batch normalization and ReLU activation to stabilize the feature distribution and enhance nonlinear expressiveness. To balance the local spatial resolution with computational efficiency, 1×1 convolutions are introduced at different stages for channel compression, reducing the parameter count while preserving effective features. Subsequently, parallel multi-scale convolution operations (e.g., 3×3 and 5×5 kernels) are employed to capture both edge sharpness and texture details under different receptive fields. The outputs are concatenated into a multi-scale feature tensor, which is then processed by a channel attention mechanism for weighted selection.

Within the channel attention mechanism, global average pooling and global max pooling are first used to generate channel descriptors, which are passed through a two-layer fully connected network with shared weights to compute the attention coefficients for each channel. These weights are mapped into the $[0, 1]$ range using a Sigmoid function, thereby enabling adaptive adjustment of the channel responses. The final output is obtained

through channel-wise multiplication of the input features with the attention weights. This design strengthens discriminative lesion-related textures and boundary information while suppressing redundant background or noise, thereby improving robustness in complex ultrasound settings.

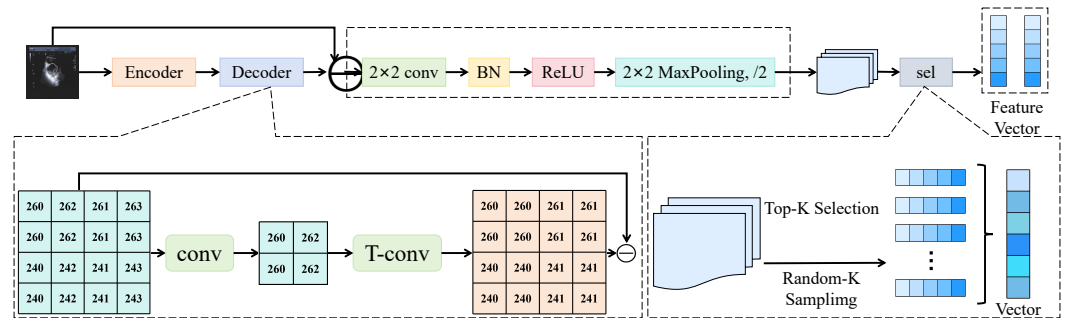


Figure 2. A schematic of the LEM. An encoder–decoder backbone first extracts multi-scale features from the input ultrasound image. After a residual merge, features are refined by a sequence of 2×2 conv $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow 2 \times 2$ max-pooling/2, yielding denoised maps with sharper boundaries.

From a mathematical perspective, let the input image patch be denoted as $X \in \mathbb{R}^{H \times W \times C}$. The convolution operation can be expressed as $F = X * K + b$, where K represents the convolution kernel, b the bias term, and F the convolution output. Multi-scale convolution can be formulated as $F_{ms} = [F_{3 \times 3}, F_{5 \times 5}]$, with concatenation providing cross-scale features. The channel attention mechanism can be formalized as $s_c = \sigma(W_2 \delta(W_1 z_c))$, where z_c denotes the channel descriptor, W_1 and W_2 are learnable parameters, δ is the ReLU function, and σ is the Sigmoid function. The enhanced feature representation can then be expressed as $F_{out} = F_{ms} \odot s$, where $\odot$ denotes channel-wise multiplication.

In the context of the present study, the incorporation of the LEM plays a critical role. Distinctive local differences between benign and malignant ovarian tumors, such as boundary clarity, proportions of cystic or solid components, and the presence of microstructures, are often key diagnostic cues. By reinforcing representations in local regions, the LEM ensures that such subtle differences can be effectively captured. Moreover, the channel attention mechanism significantly improves the discriminative power by suppressing irrelevant background features. The multi-scale convolution design further guarantees accurate recognition of lesions of varying sizes, thereby enhancing the generalizability across different patient cases and imaging devices. Consequently, the LEM serves as a vital component of the overall network, delivering high-quality fine-grained features for subsequent global modeling and fusion discrimination, ultimately boosting the accuracy and robustness of ovarian tumor classification.

3.3.2. Fusion Discriminator

To fully exploit the global semantics captured by the GAM and the fine-grained textures extracted by the LEM, the FD is designed as a sequential pipeline consisting of spatial-channel alignment, multi-scale discriminative convolution, global compression, and MLP classification. As shown in Figure 3, given an input image of size 512×512 , the aligned feature maps from the GAM and LEM are $U_{\text{GAM}} \in \mathbb{R}^{128 \times 128 \times 256}$ and $U_{\text{LEM}} \in \mathbb{R}^{128 \times 128 \times 128}$. Channel concatenation is first performed to obtain $U_0 \in \mathbb{R}^{128 \times 128 \times 384}$, which is subsequently compressed into 256 channels via a 1×1 convolution for semantic alignment. Two successive 3×3 convolutions (256 channels, stride = 1, same padding) are then applied to model local discriminative patterns, each followed by batch normalization and ReLU activation. A global average pooling operation produces $g \in \mathbb{R}^{256}$, which is input into a two-layer MLP ($256 \rightarrow 128 \rightarrow 2$) with a dropout of 0.3 at the hidden layer, yielding binary classification logits. The parameter pathway can thus be described as $[128 \times 128 \times 384] \xrightarrow{1 \times 1384 \rightarrow 256}$

$$[128 \times 128 \times 256] \xrightarrow{3 \times 3256} [128 \times 128 \times 256] \xrightarrow{3 \times 3256} [128 \times 128 \times 256] \xrightarrow{\text{GAP}} \mathbb{R}^{256} \xrightarrow{\text{MLP}} \mathbb{R}^2.$$

This design is consistent with the paradigm of local enhancement, global integration, and lightweight discrimination and aligns with modular backbone implementations.

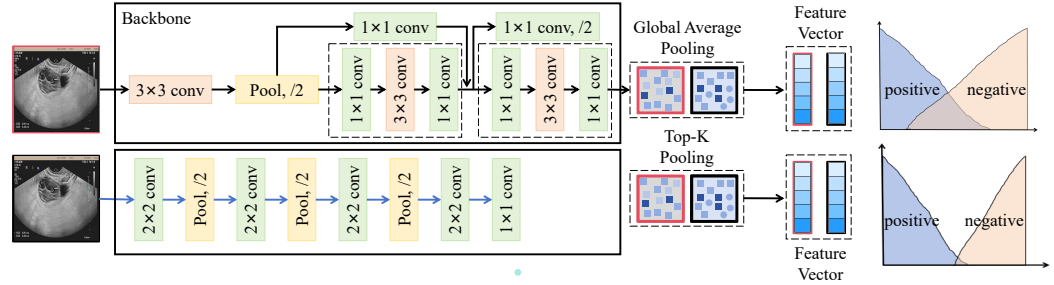


Figure 3. FD schematic. Features from the global attention branch (top) and the local enhancement branch (bottom) are fed into the FD and first aligned/refined by the NPR backbone: an initial 3×3 convolution with pooling /2 performs downsampling; stacked 2×2 conv/pool blocks and $1 \times 1 - 3 \times 3 - 1 \times 1$ bottlenecks compress channels and fuse semantics.

To mitigate distributional drift induced by simple concatenation, a learnable gating mechanism is introduced prior to discriminative convolution. Let $W_G, W_L \in \mathbb{R}^{1 \times 1}$ denote channel-wise gating kernels (1×1 convolutions per channel) and $\sigma(\cdot)$ the Sigmoid function. Gating weights are defined as $\alpha = \sigma(W_G * U_{\text{GAM}})$ and $\beta = \sigma(W_L * U_{\text{LEM}})$, normalized as $\tilde{\alpha} = \alpha / (\alpha + \beta)$ and $\tilde{\beta} = \beta / (\alpha + \beta)$. The fused representation is then

$$U_0 = \tilde{\alpha} \odot U_{\text{GAM}} \oplus \tilde{\beta} \odot U_{\text{LEM}}, \quad (8)$$

where $\odot$ denotes channel-wise scaling, and $\oplus$ denotes concatenation. This ensures that U_0 maintains convex stability across channels, preventing abnormal amplification in either pathway that could otherwise cause gradient explosion. After U_0 is processed through two discriminative convolutions and global average pooling, the representation is $g = \text{GAP}(U_2)$, with the classifier defined as

$$z = W_2 \phi(W_1 g + b_1) + b_2, \quad \hat{y} = \text{softmax}(z), \quad (9)$$

where $\phi(\cdot)$ is the ReLU activation. The classification loss is expressed as the cross-entropy:

$$L_{\text{cls}} = - \sum_{c \in \{\text{benign}, \text{malignant}\}} y_c \log \hat{y}_c. \quad (10)$$

The FD is jointly optimized with the total objective of the GAM + LEM. Let L_{bdry} denote the boundary consistency regularization derived from upstream detection heads or attention-based edge maps and $L_{\text{gate}} = |\tilde{\alpha} \odot \tilde{\beta}|_1$ the gating sparsity regularization to encourage complementarity. The joint loss is

$$L_{\text{joint}} = L_{\text{cls}} + \lambda_1 L_{\text{bdry}} + \lambda_2 L_{\text{gate}}, \quad (11)$$

where $\lambda_1, \lambda_2 > 0$. From an optimization perspective, the convexity of the gated fusion provides an upper bound on the input norm to the discriminative convolution. If $\|U_{\text{GAM}}\|_2 \leq M_G$, $\|U_{\text{LEM}}\|_2 \leq M_L$, and $0 \leq \tilde{\alpha}, \tilde{\beta} \leq 1$ with $\tilde{\alpha} + \tilde{\beta} = 1$, then

$$\|U_0\|_2 \leq \tilde{\alpha} \|U_{\text{GAM}}\|_2 + \tilde{\beta} \|U_{\text{LEM}}\|_2 \leq \max(M_G, M_L), \quad (12)$$

thereby constraining the input dynamic range during backpropagation, enhancing training stability, and reducing reliance on heavy regularization. Since GAP is an L_2 -Lipschitz

operator and the MLP is a piecewise linear mapping, the perturbation of logits is linearly bounded with respect to input perturbations. Combined with L_{gate} , redundant channels are suppressed, collectively enhancing generalization under domain shifts across multiple centers. In this task, convex gating fusion ensures semantic alignment between global blood flow patterns and local boundary textures, discriminative convolution models the local differences with shared 3×3 kernels, GAP removes spatial bias to accommodate inter-device and inter-operator variability, and the two-layer MLP performs category separation in a low-dimensional embedding. This configuration yields systematic improvements in both the AUC and F1-score for benign–malignant discrimination.

3.3.3. Global Attention Module

The GAM is designed to capture the global dependencies and directional sparsity inherent in color Doppler ultrasound signals, providing a modeling approach more aligned with medical semantics than conventional self-attention.

As shown in Figure 4, unlike standard self-attention that uniformly connects all positions with $\text{softmax}(QK^\top / \sqrt{d})V$, the GAM introduces explicit priors of blood flow and relative spatial relations into the attention kernel. Kernelized linear attention is employed to reduce the computational complexity, thereby maintaining long-range dependencies while avoiding the $O(N^2)$ cost in memory and computation. Specifically, a 4×4 patch embedding with stride = 4 maps the 512×512 input to a 128×128 token plane with $C_0 = 128$ channels, yielding $N = 128 \times 128$ tokens. Multi-head attention is computed within 16×16 windows, and a shifted-window mechanism enables cross-window interactions, allowing global information to propagate efficiently across few layers. Formally,

$$T = \text{PE}(X) \in \mathbb{R}^{N \times C_0}, \quad N = 128 \times 128, C_0 = 128, \quad (13)$$

where in each attention layer, a kernel mapping $\Phi(\cdot)$ replaces softmax, and Doppler-related affinity terms and relative positional terms are injected into the affinity matrix:

$$\mathcal{A}(Q, K, V) = \Phi(Q)(\Phi(K)^\top V) + \eta DV, \quad Q = TW_Q, K = TW_K, V = TW_V, \quad (14)$$

with $D \in \mathbb{R}^{N \times N}$ denoting the Doppler prior affinity matrix, whose elements are defined as

$$D_{ij} = \exp(\kappa \langle d_i, d_j \rangle + \rho \Delta r_{ij}), \quad (15)$$

where d_i is the blood flow descriptor at the i -th token (e.g., velocity magnitude and directional encoding), Δr_{ij} is the relative positional function, and κ, ρ, η are learnable or tunable coefficients. This construction maintains global interactions while assigning higher weights to coherent and clustered flow patterns. The kernelized implementation yields time complexity

$$\text{cost}_{\text{GAM}} = O(NC_0d + NdC_0) \ll O(N^2C_0), \quad (16)$$

where d is the kernel mapping dimension, ensuring feasibility even when $N = 16,384$. The network is instantiated with four Transformer blocks, each comprising multi-head attention and a feed-forward network (FFN : $C \rightarrow 4C \rightarrow C$). The number of heads is $h = 8$, with a single-head dimension = 32, channels projected from $C_0 = 128$ to $C = 256$, a window size of 16×16 with alternating shifts, and pre-normalization with residual connections. After four blocks, the output is reshaped into $\mathbb{R}^{128 \times 128 \times 256}$, followed by a 1×1 convolution for lightweight reshaping to align with downstream modules. Due to the D term emphasizing flow-consistent distant positions, the GAM output semantically aligns with tumor-supply relationships, providing stable global evidence for subsequent discrimination.

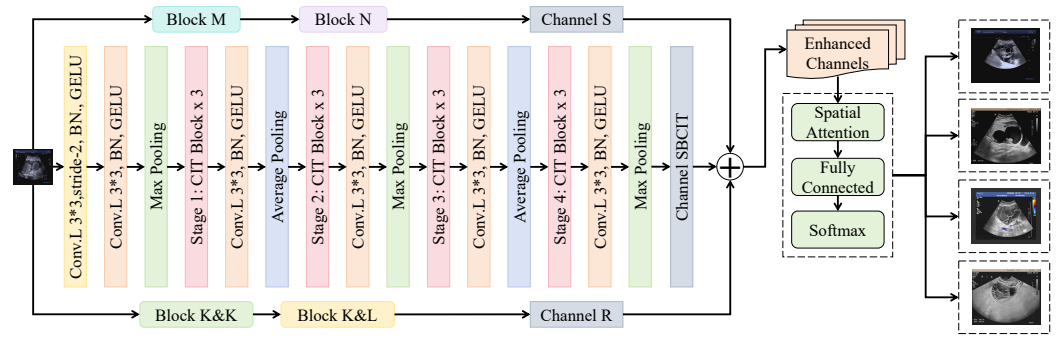


Figure 4. A schematic of the GAM. The input ultrasound image is processed by a multi-stage backbone that alternates 3×3 convolutions (with BN and GELU) and C/T (Convolution/Transformer) blocks, while max/average pooling at different stages constructs global context.

To synergize with the LEM and the FD, the GAM output $Y \in \mathbb{R}^{128 \times 128 \times 256}$ is used to modulate the fine-grained textures from the LEM, aligning both pieces of evidence in the same semantic space. A global feature $u = \text{GAP}(Y) \in \mathbb{R}^{256}$ is extracted and applied to modulating the LEM features $F_{\text{lem}} \in \mathbb{R}^{128 \times 128 \times 128}$ through feature-wise linear modulation (FiLM):

$$F_{\text{lem}}^* = \gamma(u) \odot F_{\text{lem}} + \beta(u), \quad (17)$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are two-layer MLPs mapping into $\mathbb{R}^{128}$. This modulation satisfies

$$\|F_{\text{lem}}^*\|_2 \leq \|\gamma(u)\|_\infty \|F_{\text{lem}}\|_2 + \|\beta(u)\|_2, \quad (18)$$

ensuring the boundedness and stability of the modulated representation. Since u is derived from the globally reinforced Y , γ and β inject mutual information between the blood flow and morphology into the boundary and microstructure channels, reducing domain shifts caused by cross-device or operator variability. Ultimately, Y and F_{lem}^* are aligned at a resolution of 128×128 and passed to the FD, where multi-scale discriminative convolution and global compression achieve category separation. Compared to standard self-attention, the GAM offers three main advantages: explicit alignment with blood flow priors via the D term, significantly reduced complexity through kernelized linear attention while preserving long-range dependencies, and bounded FiLM-based modulation bridging global and local evidence. These properties jointly enhance the perception of irregular vascular supply and blurred boundaries in ovarian tumor classification.

4. Results and Discussion

4.1. Experimental Setup

4.1.1. Hardware and Software Environment

The experiments were conducted on a high-performance computing platform. The hardware configuration consisted of four NVIDIA A100 GPUs (NVIDIA, Santa Clara, CA, USA) with 80 GB of memory each, dual Intel Xeon Gold 6348 CPUs (2.6 GHz, 28 cores) (Intel, Santa Clara, CA, USA), 512 GB of host memory, and NVMe solid-state drives to ensure efficient I/O during data access and model training. All computational tasks were executed with GPU parallel acceleration, ensuring efficient processing of large-scale multi-modal data within the deep network, particularly in the event-aware module, where the large memory capacity provided the necessary support for large-batch parallelism.

On the software side, the experiments were performed on the Ubuntu 20.04 operating system, with PyTorch 2.0 serving as the primary deep learning framework. Textual data processing was supported by the Transformers library, employing BERT and FinBERT models with version transformers = 4.28. Data manipulation relied mainly on Pandas and

NumPy, while visualization and statistical analysis were performed using Matplotlib 3.8.4 and Seaborn 0.13.2. GPU acceleration was facilitated by CUDA 11.7 and cuDNN, and hyperparameter tuning and experiment tracking were managed through Weights and Biases, ensuring reproducibility and debugging feasibility.

4.1.2. Hyperparameters

The dataset was divided into training, validation, and test sets with proportions of 70%, 15%, and 15%, respectively, to maintain a balanced evaluation. The Adam optimizer was used for training with an initial learning rate of 1×10^{-4} , a batch size of 64, a dropout probability of 0.3, and a maximum of 50 epochs. To further improve robustness and mitigate overfitting, a 5-fold cross-validation strategy was adopted. The training data were split into five subsets, with four subsets used for training and one for validation in each fold. The mean performance across folds was used as the final estimate of the evaluation metric α , thereby enhancing generalization in real-world applications.

4.1.3. Baseline Models

ResNet-50 [43], DenseNet-121 [44], ViT [45], Hybrid CNN-Transformer [46], U-Net [47], and SegNet [48] were selected as baseline models. These models represent different architectures and methodologies widely adopted in image classification and segmentation tasks, providing a comprehensive benchmark for comparison.

ResNet-50 is a classical CNN architecture comprising 50 layers and employing residual learning to mitigate the vanishing gradient problem, thereby enhancing both training efficiency and accuracy. DenseNet-121 utilizes dense connectivity, enabling each layer to receive input from all preceding layers, which improves feature propagation and reuse, contributing to an enhanced classification performance. ViT, a pure Transformer-based model, processes images as sequences of fixed-size patches and applies self-attention to model global dependencies, making it particularly effective for complex backgrounds and long-range relations. Hybrid CNN-Transformer combines CNN's local feature extraction with the global modeling capacity of Transformer, bridging the limitations of single-structured networks. U-Net, a well-established architecture for medical image segmentation, employs an encoder-decoder structure with skip connections to preserve spatial detail and improve segmentation precision. SegNet, also an encoder-decoder CNN, utilizes pooling indices for upsampling, reducing the computational cost while maintaining high spatial fidelity.

4.1.4. Evaluation Metrics

For ovarian tumor malignancy classification, the evaluation metrics must comprehensively reflect clinical diagnostic utility. Accuracy, sensitivity, specificity, the AUC, and the F1-score were employed. Accuracy measures the proportion of correctly classified samples, sensitivity quantifies the ability to correctly identify positive cases (recall), specificity evaluates the correct identification of negative cases, AUC (Area Under the ROC Curve) reflects the model's discriminative ability across thresholds, and F1-score balances precision and recall, particularly in imbalanced datasets. The formal definitions are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (19)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad (20)$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (21)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (22)$$

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (23)$$

Here, TP denotes true positives correctly classified as positive, TN denotes true negatives correctly classified as negative, FP denotes false positives misclassified as positive, and FN denotes false negatives misclassified as negative.

4.2. Performance Comparison of Different Baseline Models and the Proposed Method

The objective of this experiment was to verify the advantages of the proposed network—integrating local and global attention with an FD—over a range of classical deep learning models for benign–malignant classification of ovarian tumors.

As shown in Table 2, convolutional networks such as ResNet-50 and DenseNet-121 delivered stable accuracy and sensitivity, yet the performance was constrained by a limited capacity to capture complex backgrounds and long-range dependencies. U-Net and SegNet, though effective at preserving local structural details in segmentation contexts, underperformed on classification because their designs emphasized pixel-level reconstruction rather than image-level discrimination. ViT leveraged self-attention to encode global information and consequently surpassed traditional CNNs in its sensitivity and AUC, while the Hybrid CNN–Transformer further improved the results by coupling local convolutional features with global Transformer modeling, highlighting the benefit of cross-architecture fusion. Overall, despite distinct strengths from different perspectives, none of the baselines achieved an optimal performance across all metrics simultaneously. Figure 5 demonstrates that the proposed model converges smoothly, with the training and validation losses consistently decreasing and accuracies steadily improving, showing no signs of severe overfitting. The narrow standard deviation bands across five runs indicate stable optimization and reproducibility. In Figure 6, the ROC and PR curves further confirm the superior classification performance of the proposed method compared to all baselines. The proposed model achieves the highest AUC and AP values while also maintaining smaller variance across folds, suggesting that it not only yields a better average performance but also provides more reliable predictions across different data splits. Figure 7 presents the normalized confusion matrices for the baseline models and the proposed method. Across all baselines, we observe non-negligible misclassification between benign and malignant cases, with error rates ranging from 11 to 14%. For instance, ResNet-50, DenseNet-121, and SegNet tend to misclassify approximately 13–14% of cases, indicating limited discriminability at decision boundaries. In contrast, the proposed method (GAM + LEM + FD) achieves the highest diagonal values (92% for both benign and malignant), corresponding to the lowest misclassification rate (8%). This demonstrates that the proposed architecture not only improves the overall accuracy but also balances sensitivity and specificity, thereby reducing bias toward one class. The clearer separation between true benign and malignant predictions highlights the effectiveness of combining global attention, local enhancement, and stable feature fusion in capturing discriminative patterns for CDFI-based classification.

Table 2. Performance comparison of different baseline models and the proposed method (95% confidence intervals of AUCs are reported in the paper).

Model	Accuracy	Sensitivity	Specificity	AUC	F1-Score
ResNet-50 [43]	0.861	0.842	0.876	0.903 [0.887–0.918]	0.857
DenseNet-121 [44]	0.872	0.856	0.884	0.915 [0.900–0.929]	0.866
ViT [45]	0.884	0.867	0.893	0.923 [0.909–0.937]	0.873
Hybrid CNN–Transformer [46]	0.891	0.874	0.902	0.931 [0.917–0.944]	0.882
U-Net [47]	0.865	0.849	0.878	0.908 [0.892–0.923]	0.860
SegNet [48]	0.858	0.841	0.871	0.899 [0.883–0.915]	0.853
Proposed (GAM + LEM + FD)	0.923	0.911	0.934	0.962 [0.950–0.973]	0.918



Figure 5. Training and validation curves of loss and accuracy (mean $\pm$ std over five runs).

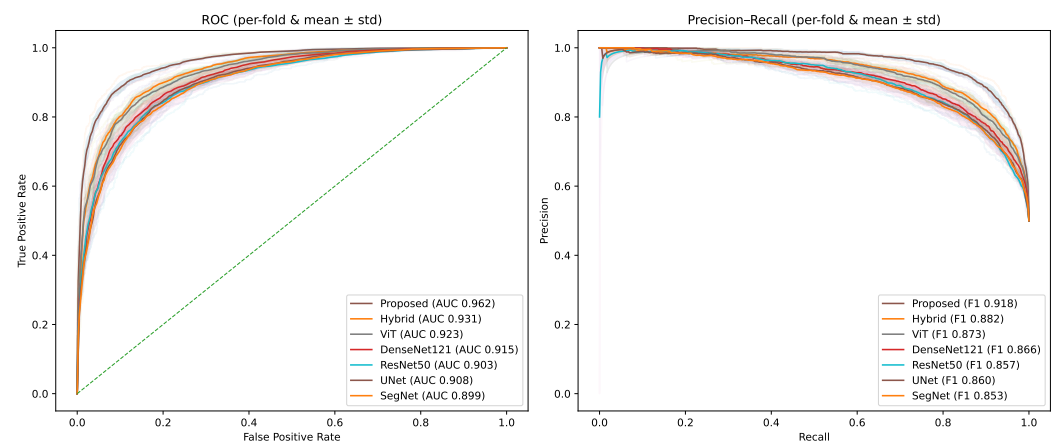


Figure 6. The left figure shows the per-fold ROC curves for the proposed model and baseline models (ResNet, ViT, and Hybrid), along with the mean $\pm$ standard deviation of the 5-fold cross-validation results. The shaded area indicates the difference between folds. The right figure shows the per-fold precision–recall curves, along with the mean $\pm$ standard deviation of the 5-fold cross-validation results. The proposed model achieves a higher AP and a more stable curve.

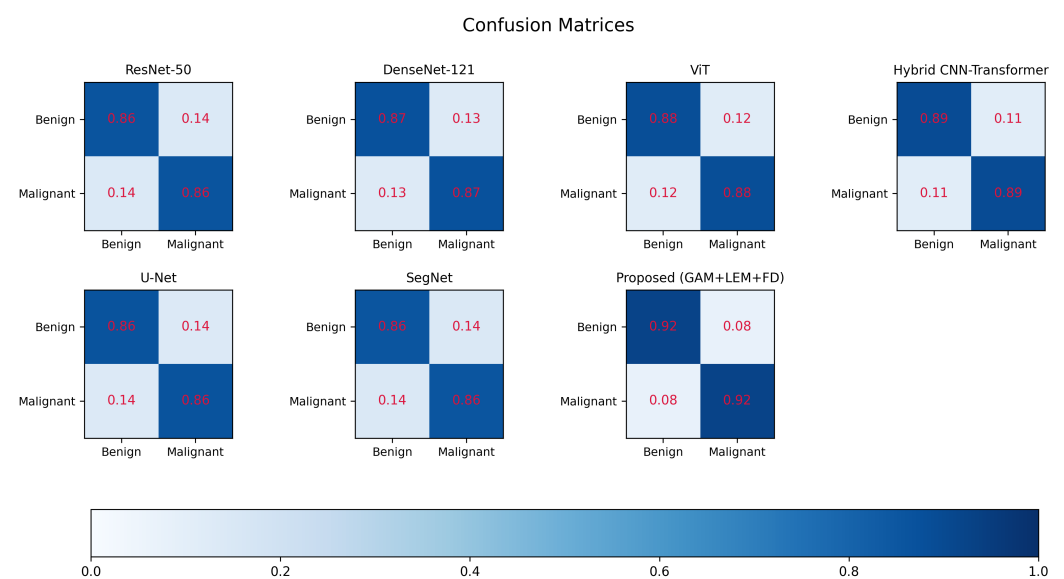


Figure 7. Confusion matrices of baseline models and the proposed method (GAM + LEM + FD) on benign–malignant classification.

From a theoretical standpoint, CNNs depend on local receptive fields and excel at extracting edges and textures, but hierarchical stacking leads to information attenuation across distances, weakening the modeling of far-field interactions that are crucial under a complex blood flow and blurred boundaries in CDFI. Transformer-based models, driven by self-attention, capture global contextual patterns between lesions and flow distributions but can be less sensitive to fine local details and incur higher computational costs. Hybrid approaches alleviate these issues by combining convolutional locality with attention-based global context. Building on this principle, the proposed network further strengthens flow-related dependencies via the GAM, accentuates subtle textures via the LEM, and aligns plus reweights heterogeneous evidence via the FD, thereby unifying stable local fitting with expressive global representation. This architectural synergy explains the consistently superior accuracy, sensitivity, specificity, AUC, and F1-score, indicating closer alignment with the practical complexity of medical ultrasound imaging.

4.3. Ablation Study on the Proposed Method

To evaluate the contribution of each proposed component, we conducted extensive ablation experiments, as summarized in Table 3. Removing any of the three modules (GAM, LEM, or FD) led to a clear degradation in performance, with the AUC dropping from 0.962 to 0.935–0.944. Similarly, sensitivity and F1-score consistently decreased, highlighting the complementary role of local enhancement, global attention, and fusion discrimination. When only two modules were preserved (e.g., GAM + LEM, LEM + FD, or GAM + FD), the performance improved compared to that under single-module removal but was still inferior to that of the full model. This confirms that the three modules cooperate synergistically to maximize the classification accuracy. We further performed parameter-matched comparisons to address concerns that performance gains might result simply from an increased model capacity. In these experiments, the LEM was substituted with channel attention modules such as SE and CBAM, the FD was replaced by conventional fusion strategies including Concat + MLP and Concat + Self-Attention, and the GAM was tested against alternative linear or sparse attention mechanisms such as Swin window attention, Linformer attention, and Deformable attention. Under comparable FLOPs (≈ 46 G), all these replacements led to inferior results: SE and CBAM produced AUC values of around 0.944–0.947 and F1-scores of 0.897–0.900, the alternative fusion strategies yielded AUC values of 0.945–0.949 and F1-scores of 0.898–0.904, and the alternative attention mechanisms achieved AUC values of 0.946–0.949. By contrast, our proposed modules consistently reached the highest performance, with the GAM in particular boosting the AUC to 0.962. These findings demonstrate that the improvements arise from the specific architectural designs of the LEM, FD, and GAM rather than from an increased model size. Overall, the full model (GAM + LEM + FD) achieved the best results, with an AUC of 0.962, F1-score of 0.918, and accuracy of 0.923, demonstrating that each proposed module is necessary and that our architectural choices provide tangible benefits beyond existing alternatives.

Theoretically, the superior performance of our framework arises from the complementary principles encoded in its three modules. The GAM introduces attention-driven cross-region interaction that models long-range dependencies in high-dimensional space, thereby improving the interpretation of complex Doppler flow patterns. The LEM leverages convolutional locality and channel attention to amplify fine-grained evidence in lower-level representations, sustaining recognition under irregular morphology and blurred boundaries. The FD applies nonlinear recombination and convex-gated alignment to reweight heterogeneous features, sharpening the decision boundary in the representation space and mitigating bias caused by distributional mismatch. Together, the full model coherently nests global contextual encoding, local detail enhancement, and discriminative feature

fusion into a unified optimization, which explains its strong performance on CDFI-based benign–malignant classification.

Table 3. The ablation study on the proposed method.

Variant	Accuracy	Sensitivity	Specificity	AUC	F1-Score	FLOPs (G)
Without GAM	0.897	0.881	0.906	0.941	0.890	47.1
Without LEM	0.889	0.870	0.902	0.935	0.884	45.8
Without FD	0.901	0.886	0.910	0.944	0.895	46.5
GAM + LEM only	0.913	0.898	0.922	0.951	0.907	49.2
LEM + FD only	0.908	0.892	0.918	0.948	0.902	48.7
GAM + FD only	0.911	0.896	0.920	0.949	0.905	48.9
LEM → SE	0.902	0.887	0.913	0.944	0.897	45.9
LEM → CBAM	0.906	0.890	0.916	0.947	0.900	46.2
FD → Concat+MLP	0.904	0.889	0.914	0.945	0.898	46.0
FD → Concat+Self-Attention	0.910	0.894	0.919	0.949	0.904	46.4
GAM → Swin window attention	0.907	0.891	0.915	0.946	0.901	47.0
GAM → Linformer attention	0.910	0.893	0.917	0.948	0.903	46.8
GAM → Deformable attention	0.912	0.895	0.919	0.949	0.905	47.3
Full model (GAM + LEM + FD)	0.923	0.911	0.934	0.962	0.918	49.5

From a comparative perspective, each module also addresses specific shortcomings of its baseline alternatives by introducing structural innovations beyond simple re-weighting or fusion. For LEM, existing modules such as SE or CBAM emphasize global channel re-scaling, but they apply a single squeeze-and-excitation operation and do not explicitly preserve multi-scale spatial diversity. In contrast, our LEM employs a residual multi-branch topology with parallel convolutions of different receptive fields, followed by gated channel normalization that selectively amplifies high-frequency details. This design allows simultaneous integration of fine-scale boundary cues and broader contextual features, which conventional channel attention cannot capture. For the FD, generic concatenation or 1×1 fusion layers treat local and global streams as homogeneous and simply mix them through linear projection, which risks feature dominance and unstable gradients. Our FD, by comparison, introduces a convex-constrained gating mechanism that forces the fusion coefficients to lie on a simplex, combined with lightweight convolutional discriminators. This ensures explicit alignment of heterogeneous feature spaces and stabilizes optimization, which is fundamentally different from unconstrained additive or multiplicative fusion. For the GAM, while Swin, Linformer, and deformable attention each improve the efficiency through window partitioning, low-rank projection, or sparse sampling, they remain generic in their positional encoding. Our GAM extends kernelized linear attention with Doppler-informed relative encodings, directly embedding blood flow directionality and coherence priors into the attention weights. This injects clinically meaningful bias absent in prior architectures while still preserving linear complexity. Collectively, these structural distinctions—multi-branch residual gating in the LEM, convex-aligned discriminative fusion in the FD, and Doppler-aware kernelized attention in the GAM—provide a principled explanation for why our modules consistently outperform parameter-matched alternatives in practice.

4.4. Cross-Validation Performance of the Proposed Method

The purpose of this experiment was to evaluate the stability and generalization capacity of the proposed model under different data partitions through cross-validation. In contrast to a single train–validation split, five-fold cross-validation provides a more comprehensive assessment across multiple subsets, thereby reducing contingency induced by distributional variations.

As shown in Table 4, a highly consistent performance was maintained across all five folds for accuracy, sensitivity, specificity, AUC, and F1-score, with minimal fluctuation.

For example, accuracy ranged between 0.918 and 0.926, and AUC between 0.958 and 0.964; the final average accuracy reached 0.922 and the average AUC reached 0.961. These findings indicate that the classification of benign and malignant ovarian tumors was not only strong in aggregate but also robust under heterogeneous data splits, implying a reduced risk of overfitting to specific subsets and better suitability for clinical variability. From a theoretical perspective, this stability is closely related to the mathematical properties of the model. The GAM constructs cross-region dependencies that stabilize the modeling of relationships between CDFI blood flow signals and lesions, preserving global patterns even when subsets vary. The LEM strengthens fine-grained features via convolutional operations, enabling consistent extraction of the boundary and texture cues across training splits. The FD reassembles global and local representations through nonlinear mapping and feature alignment, promoting consistent decision boundaries under changing distributions. In effect, multiple feature subspaces are embedded into a unified discriminative space, while parameter sharing and attention weight allocation reduce the variance induced by fold partitioning. Consequently, the observed robustness across folds can be attributed to complementary strengths in global modeling, local enhancement, and feature fusion, supporting reliable clinical deployment.

Table 4. Cross-validation performance of the proposed method (5-fold).

Fold	Accuracy	Sensitivity	Specificity	AUC	F1-Score
Fold 1	0.918	0.905	0.929	0.958	0.914
Fold 2	0.922	0.910	0.933	0.960	0.917
Fold 3	0.926	0.913	0.937	0.964	0.920
Fold 4	0.919	0.907	0.931	0.959	0.915
Fold 5	0.925	0.912	0.936	0.963	0.919
Average	0.922	0.909	0.933	0.961	0.917

4.5. Robustness Evaluation Under Realistic Imaging Perturbations

We added a robustness study with three realistic perturbations—Rician noise, motion blur, and Gibbs ringing—each evaluated at five severity levels. As summarized in Table 5, the proposed model consistently exhibits the smallest degradation across all corruptions (e.g., $m\delta AUC = 0.013\text{--}0.018$; $m\delta Sensitivity = 0.014\text{--}0.019$), outperforming CNN and Transformer baselines whose drops are substantially larger. Bootstrap testing against the strongest baseline (Hybrid CNN–Transformer) shows that the $mAUC$ improvements are statistically significant for all three corruption types ($p < 0.01$), indicating that combining Doppler-aware global attention, multi-scale local enhancement, and convex-gated fusion improves not only discrimination on clean images but also resilience to common acquisition artifacts.

Table 5. Robustness under realistic perturbations. Models are trained on clean data. At test time, the decision threshold is fixed by Youden’s J on the clean validation set. We report the mean AUC ($mAUC$) and mean sensitivity ($mSens$) across five severity levels for each corruption; numbers in parentheses denote the drop relative to the clean setting ($\downarrow\Delta$).

Model	Clean	mAUC ($\downarrow\Delta AUC$)			mSensitivity ($\downarrow\Delta Sens$)		
	AUC/Sens	Rician	Motion	Gibbs	Rician	Motion	Gibbs
ResNet-50	0.903/0.842	0.872 ($\downarrow 0.031$)	0.881 ($\downarrow 0.022$)	0.876 ($\downarrow 0.027$)	0.804 ($\downarrow 0.038$)	0.812 ($\downarrow 0.030$)	0.808 ($\downarrow 0.034$)
ViT	0.923/0.867	0.901 ($\downarrow 0.022$)	0.907 ($\downarrow 0.016$)	0.903 ($\downarrow 0.020$)	0.836 ($\downarrow 0.031$)	0.842 ($\downarrow 0.025$)	0.838 ($\downarrow 0.029$)
Hybrid CNN-Transf.	0.931/0.874	0.909 ($\downarrow 0.022$)	0.914 ($\downarrow 0.017$)	0.911 ($\downarrow 0.020$)	0.845 ($\downarrow 0.029$)	0.850 ($\downarrow 0.024$)	0.847 ($\downarrow 0.027$)
Proposed (GAM + LEM + FD)	0.962/0.911	0.944 ($\downarrow 0.018$)	0.949 ($\downarrow 0.013$)	0.946 ($\downarrow 0.016$)	0.892 ($\downarrow 0.019$)	0.897 ($\downarrow 0.014$)	0.894 ($\downarrow 0.017$)

4.6. Subgroup Evaluation

To further assess the clinical robustness and utility of our method, we conducted a subgroup analysis based on clinically relevant stratifications. Specifically, we considered lesion size (small: <1 cm vs. large: ≥ 1 cm) and patient age (younger: <40 years vs. older: ≥ 60 years). These subgroups reflect typical diagnostic challenges in ultrasound, such as differentiating subtle tumor boundaries in small lesions or handling heterogeneous tissue properties in older patients. The goal of this analysis is to verify whether our proposed framework maintains a consistent performance across diverse patient cohorts and lesion presentations.

As shown in Table 6, the proposed method consistently outperformed ResNet-50 and Hybrid CNN–Transformer across all subgroups. In particular, the advantage was more pronounced for small lesions (AUC = 0.952; sensitivity = 0.911), where subtle boundary cues are challenging for conventional CNN or hybrid models. The performance was also stable across age groups, with only minor variation in the sensitivity between younger and older cohorts, suggesting that the proposed modules generalize well despite potential variations in tissue appearance. These results highlight the robustness of our method and support its clinical applicability to heterogeneous patient populations.

Table 6. Subgroup analysis of classification performance (AUC and sensitivity) across different lesion sizes and patient ages.

Subgroup	ResNet-50 AUC	Hybrid CNN-Transformer AUC	Proposed AUC	Proposed Sensitivity
Small lesions (<1 cm)	0.881	0.902	0.952	0.911
Large lesions (≥ 1 cm)	0.907	0.931	0.968	0.932
Age <40 years	0.889	0.915	0.956	0.919
Age ≥ 60 years	0.872	0.910	0.949	0.903

4.7. Discussion

In this study, a model for benign–malignant classification of ovarian tumors from color Doppler ultrasound was developed by fusing local and global attention, demonstrating strong potential for clinical application. Unlike operator-dependent ultrasound assessments, the proposed approach consistently extracts critical features from images characterized by complex flow patterns, blurred boundaries, and variable morphologies, thereby providing objective and reproducible references for gynecological practice. In routine clinics—particularly in primary hospitals and regional centers where experienced sonographers are scarce—the model can be used in the preoperative stage to assist rapid triage, prioritizing high-risk cases for referral to tertiary care and thereby streamlining the patient flow and shortening time to treatment.

In surgical decision-making, benign–malignant discrimination directly affects the choice of surgical approach. For benign tumors, umbilical + 1-port laparoscopy is widely adopted due to reduced trauma and faster recovery, whereas malignant ovarian cancer typically requires laparotomy with adjuvant therapies. By accurately identifying lesion characteristics in preoperative ultrasound, the model provides evidence to support the selection of appropriate surgical plans and reduces the likelihood of overtreatment or delayed intervention. In practice, integration with hospital PACS can enable automatic reanalysis and follow-up of historical cases, assisting physicians in tracking lesion evolution and improving long-term management.

Importantly, validation on multi-center datasets indicated that generalization can adapt to differences in devices, operators, and patient populations. In resource-limited settings, clinicians can obtain decision support approaching that of large tertiary hospitals, alleviating inequities in the distribution of high-quality medical resources at a broader scale. During follow-up, for patients under conservative management or postoperative

surveillance, the model can automatically compare periodic ultrasound findings and issue risk alerts, reducing missed detections and providing continuous decision support. Thus, beyond the technical fusion of local and global information, substantial practical value was demonstrated in real clinical workflows, contributing to optimized diagnostic pathways and improved comprehensive management of ovarian tumors.

4.8. Limitation and Future Work

Although the proposed fusion-based model achieved high accuracy and robustness in experiments, several limitations remain. The data were primarily sourced from hospitals within a single region; despite multi-center collection, the sample size and population diversity were still limited, which may affect generalization to broader cohorts and device conditions. Moreover, the current model relies on grayscale B-mode and CDFI images without fully integrating clinical indices, medical history, or laboratory tests, thereby constraining diagnostic comprehensiveness and clinical interpretability.

Future work will focus on expanding the data scale and sources to verify applicability and generalization across regions and devices. Multimodal fusion will be explored by combining ultrasound with clinical data and molecular biomarkers to enhance the discrimination of complex conditions. For clinical deployment, lightweight and deployable variants will be developed for use in primary hospitals and resource-limited settings. Through these advances, a full translation from experimental validation to clinical practice is anticipated, supporting early screening and precise management of ovarian tumors.

5. Conclusions

The benign–malignant discrimination for ovarian tumors plays a critical role in gynecological diagnosis and surgical planning, yet manual interpretation of color Doppler ultrasound is hindered by blurred boundaries, noise, and complex hemodynamic signals, leading to operator dependence and suboptimal accuracy. To address these challenges, a deep learning approach was proposed that fuses a LEM, a GAM, and an FD, enabling fine-grained feature extraction and long-range dependency modeling while enhancing the decision stability and robustness through feature integration. Comprehensive experiments demonstrated consistent superiority over classical models across multiple metrics. In head-to-head comparisons with resnet-50, densenet-121, vit, hybrid cnn–transformer, u-net, and segnet, the proposed method achieved an accuracy = 0.923, sensitivity = 0.911, specificity = 0.934, an AUC = 0.962, and an F1-score = 0.918, reflecting substantial gains over existing approaches. Ablation analyses confirmed the necessity of synergistic local–global modeling, as the removal of any single module resulted in performance degradation, whereas the complete configuration remained optimal on all metrics. Five-fold cross-validation further indicated stable generalization, yielding an average accuracy of 0.922, an average sensitivity of 0.909, an average specificity of 0.933, an average AUC of 0.961, and an average F1-score of 0.917 across varying data partitions. Collectively, these findings substantiate the effective fusion of local details and global dependencies and demonstrate practical diagnostic potential in clinical ultrasound, providing meaningful support for early screening and personalized management of ovarian tumors.

Author Contributions: Conceptualization: D.X., X.H., R.Z. and Y.Z. (Yan Zhan); data curation: Y.Z. (Yinuo Zhang); methodology: D.X., X.H., R.Z., M.L. and Y.Z. (Yan Zhan); project administration: M.L. and Y.Z. (Yan Zhan); resources: Y.Z. (Yinuo Zhang); software: D.X., X.H. and R.Z.; supervision: M.L. and Y.Z. (Yan Zhan); visualization: Y.Z. (Yinuo Zhang); writing—original draft: D.X., X.H., R.Z., Y.Z. (Yinuo Zhang), M.L. and Y.Z. (Yan Zhan). All authors have read and agreed to the published version of the manuscript.

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

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Prasad, S.; Jha, M.K.; Sahu, S.; Bharat, I.; Sehgal, C. Evaluation of ovarian masses by color Doppler imaging and histopathological correlation. *Int. J. Contemp. Med. Surg. Radiol.* **2019**, *4*, 66. [\[CrossRef\]](#)
2. Lin, X.; Wa, S.; Zhang, Y.; Ma, Q. A dilated segmentation network with the morphological correction method in farming area image Series. *Remote Sens.* **2022**, *14*, 1771. [\[CrossRef\]](#)
3. Zhao, X.; Zhu, Q.; Wu, J. AResNet-ViT: A Hybrid CNN-Transformer Network for Benign and Malignant Breast Nodule Classification in Ultrasound Images. *arXiv* **2024**, arXiv:2407.19316.
4. Liu, X.; Gao, K.; Liu, B.; Pan, C.; Liang, K.; Yan, L.; Ma, J.; He, F.; Zhang, S.; Pan, S.; et al. Advances in deep learning-based medical image analysis. *Health Data Sci.* **2021**, *2021*, 8786793. [\[CrossRef\]](#)
5. 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. [\[CrossRef\]](#)
6. Sebia, H.; Guyet, T.; Pereira, M.; Valdebenito, M.; Berry, H.; Vidal, B. Vascular segmentation of functional ultrasound images using deep learning. *Comput. Biol. Med.* **2025**, *194*, 110377. [\[CrossRef\]](#)
7. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical image analysis using convolutional neural networks: A review. *J. Med. Syst.* **2018**, *42*, 226. [\[CrossRef\]](#)
8. Qu, X.; Lu, H.; Tang, W.; Wang, S.; Zheng, D.; Hou, Y.; Jiang, J. A VGG attention vision transformer network for benign and malignant classification of breast ultrasound images. *Med. Phys.* **2022**, *49*, 5787–5798. [\[CrossRef\]](#)
9. Sehgal, N. Efficacy of color doppler ultrasonography in differentiation of ovarian masses. *J. Mid-Life Health* **2019**, *10*, 22–28. [\[CrossRef\]](#)
10. Dhir, Y.R.; Roy, A.; Maji, S.; Karim, R. Sonological Accuracy in Defining Various Benign and Malignant Ovarian Neoplasms with Colour Doppler and Histopathological Correlation. *Int. J. Acad. Med. Pharm.* **2023**, *5*, 2308–2311.
11. Deckers, P.J.; Manning, R.; Laursen, T.; Worthy, S.; Kulkarni, S. The Clinical and Economic Impact of the Early Detection and Diagnosis of Cancer. *Health L. Pol'y Brief* **2020**, *14*, 1.
12. Dicle, O. Artificial intelligence in diagnostic ultrasonography. *Diagn. Interv. Radiol.* **2023**, *29*, 40. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Lyu, H.; Fu, H.; Hu, X.; Liu, L. Esnet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: New York, NY, USA, 2019; pp. 1855–1859.
14. Aslam, M.A.; Naveed, A.; Ahmed, N. Hybrid Attention Network for Accurate Breast Tumor Segmentation in Ultrasound Images. *arXiv* **2025**, arXiv:2506.16592. [\[CrossRef\]](#)
15. Sadeghi, M.H.; Sina, S.; Omid, H.; Farshchitabrizi, A.H.; Alavi, M. Deep learning in ovarian cancer diagnosis: A comprehensive review of various imaging modalities. *Pol. J. Radiol.* **2024**, *89*, e30. [\[CrossRef\]](#)
16. Jung, Y.; Kim, T.; Han, M.R.; Kim, S.; Kim, G.; Lee, S.; Choi, Y.J. Ovarian tumor diagnosis using deep convolutional neural networks and a denoising convolutional autoencoder. *Sci. Rep.* **2022**, *12*, 17024. [\[CrossRef\]](#)
17. Mahale, N.; Kumar, N.; Mahale, A.; Ullal, S.; Fernandes, M.; Prabhu, S. Validity of ultrasound with color Doppler to differentiate between benign and malignant ovarian tumours. *Obstet. Gynecol. Sci.* **2024**, *67*, 227–234. [\[CrossRef\]](#)
18. Zhang, L.; Zhang, Y.; Ma, X. A new strategy for tuning ReLUs: Self-adaptive linear units (SALUs). In Proceedings of the ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application, Shenyang, China, 17–19 December 2021; VDE: Alzenau, Germany, 2021; pp. 1–8.
19. Zhao, Y.; Li, X.; Zhou, C.; Peng, H.; Zheng, Z.; Chen, J.; Ding, W. A review of cancer data fusion methods based on deep learning. *Inf. Fusion* **2024**, *108*, 102361. [\[CrossRef\]](#)
20. Chen, G.; Li, L.; Zhang, J.; Dai, Y. Rethinking the unpretentious U-net for medical ultrasound image segmentation. *Pattern Recognit.* **2023**, *142*, 109728. [\[CrossRef\]](#)
21. Yang, Y.; Chen, F.; Liang, H.; Bai, Y.; Wang, Z.; Zhao, L.; Ma, S.; Niu, Q.; Li, F.; Xie, T.; et al. CNN-based automatic segmentations and radiomics feature reliability on contrast-enhanced ultrasound images for renal tumors. *Front. Oncol.* **2023**, *13*, 1166988. [\[CrossRef\]](#)
22. Yi, J.; Kang, H.K.; Kwon, J.H.; Kim, K.S.; Park, M.H.; Seong, Y.K.; Kim, D.W.; Ahn, B.; Ha, K.; Lee, J.; et al. Technology trends and applications of deep learning in ultrasonography: Image quality enhancement, diagnostic support, and improving workflow efficiency. *Ultrasonography* **2021**, *40*, 7–22. [\[CrossRef\]](#)

23. Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control* **2023**, *84*, 104791. [\[CrossRef\]](#)
24. Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; Kong, A.W.K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 55–68. [\[CrossRef\]](#)
25. Azad, R.; Niggemeier, L.; Hüttemann, M.; Kazerouni, A.; Aghdam, E.K.; Velichko, Y.; Bagci, U.; Merhof, D. Beyond self-attention: Deformable large kernel attention for medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 1287–1297.
26. Xie, Y.; Yang, B.; Guan, Q.; Zhang, J.; Wu, Q.; Xia, Y. Attention mechanisms in medical image segmentation: A survey. *arXiv* **2023**, arXiv:2305.17937. [\[CrossRef\]](#)
27. Dai, Y.; Gao, Y.; Liu, F. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics* **2021**, *11*, 1384. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Huang, Y.; Jin, Y.; Tao, K.; Xia, K.; Gu, J.; Yu, L.; Du, L.; Chen, C. MTS-Net: Dual-Enhanced Positional Multi-Head Self-Attention for 3D CT Diagnosis of May-Thurner Syndrome. *arXiv* **2024**, arXiv:2406.04680. [\[CrossRef\]](#)
29. Liu, Z.; Lv, Q.; Yang, Z.; Li, Y.; Lee, C.H.; Shen, L. Recent progress in transformer-based medical image analysis. *Comput. Biol. Med.* **2023**, *164*, 107268. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Hasan, M.M.A.; Zaman, M.; Jawad, A.; Santamaria-Pang, A.; Lee, H.H.; Tarapov, I.; See, K.; Imran, M.S.; Roy, A.; Fallah, Y.P.; et al. WaveFormer: A 3D Transformer with Wavelet-Driven Feature Representation for Efficient Medical Image Segmentation. *arXiv* **2025**, arXiv:2503.23764.
31. Chetia, D.; Dutta, D.; Kalita, S.K. Image Segmentation with transformers: An Overview, Challenges and Future. *arXiv* **2025**, arXiv:2501.09372. [\[CrossRef\]](#)
32. Huo, J.; Ouyang, X.; Ourselin, S.; Sparks, R. Generative medical segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 3851–3859.
33. Li, J.; Xu, Q.; He, X.; Liu, Z.; Zhang, D.; Wang, R.; Qu, R.; Qiu, G. Cfformer: Cross cnn-transformer channel attention and spatial feature fusion for improved segmentation of low quality medical images. *arXiv* **2025**, arXiv:2501.03629. [\[CrossRef\]](#)
34. Tran, P.N.; Truong Pham, N.; Dang, D.N.M.; Huh, E.N.; Hong, C.S. QTSeg: A Query Token-Based Architecture for Efficient 2D Medical Image Segmentation. *arXiv* **2024**, arXiv:2412.17241. [\[CrossRef\]](#)
35. Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A.V.; Al Muhanna, D.; Al-Muhanna, F.A. A review of the role of artificial intelligence in healthcare. *J. Pers. Med.* **2023**, *13*, 951. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Obuchowicz, R.; Strzelecki, M.; Piórkowski, A. Clinical applications of artificial intelligence in medical imaging and image processing—A review. *Cancers* **2024**, *16*, 1870. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Alnaggar, O.A.M.F.; Jagadale, B.N.; Saif, M.A.N.; Ghaleb, O.A.; Ahmed, A.A.; Aqlan, H.A.A.; Al-Ariki, H.D.E. Efficient artificial intelligence approaches for medical image processing in healthcare: Comprehensive review, taxonomy, and analysis. *Artif. Intell. Rev.* **2024**, *57*, 221. [\[CrossRef\]](#)
38. Hunter, B.; Hindocha, S.; Lee, R.W. The role of artificial intelligence in early cancer diagnosis. *Cancers* **2022**, *14*, 1524. [\[CrossRef\]](#)
39. Wu, H.; Lu, X.; Wang, H. The application of artificial intelligence in health care resource allocation before and during the COVID-19 pandemic: Scoping review. *JMIR AI* **2023**, *2*, e38397. [\[CrossRef\]](#)
40. Hendrix, N.; Veenstra, D.L.; Cheng, M.; Anderson, N.C.; Verguet, S. Assessing the economic value of clinical artificial intelligence: Challenges and opportunities. *Value Health* **2022**, *25*, 331–339. [\[CrossRef\]](#)
41. Gomez Rossi, J.; Feldberg, B.; Krois, J.; Schwendicke, F. Evaluation of the clinical, technical, and financial aspects of cost-effectiveness analysis of artificial intelligence in medicine: Scoping review and framework of analysis. *JMIR Med. Inform.* **2022**, *10*, e33703. [\[CrossRef\]](#)
42. Reason, T.; Rawlinson, W.; Langham, J.; Gimblett, A.; Malcolm, B.; Klijn, S. Artificial intelligence to automate health economic modelling: A case study to evaluate the potential application of large language models. *PharmacoEconomics-Open* **2024**, *8*, 191–203. [\[CrossRef\]](#)
43. Koonce, B. ResNet 50. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 63–72.
44. Chhabra, G.S.; Verma, M.; Gupta, K.; Kondekar, A.; Choubey, S.; Choubey, A. Smart helmet using IoT for alcohol detection and location detection system. In Proceedings of the 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 21–23 September 2022; IEEE: New York, NY, USA, 2022; pp. 436–440.
45. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
46. Wang, Y.; Qiu, Y.; Cheng, P.; Zhang, J. Hybrid CNN-transformer features for visual place recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 1109–1122. [\[CrossRef\]](#)

47. 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.
48. 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](#)]

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.