Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.075316

REVIEW

Attention-Based Medical Image Analysis: Architectures, Applications, and Future Directions

Xinjie Yao[1], Junjie Zhu[2], Tao Hong[3,4], Dengyu Zhao[5], Weikai Liu[6] and Guangsheng Xie[7,*]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China
[2]School of Artificial Intelligence, Xiangyang Polytechnic University, Xiangyang, China
[3]China Nuclear Power Engineering Co., Ltd., Beijing, China
[4]CNNC Engineering Research Center for Fuel Reprocessing, Beijing, China
[5]School of Mathematics, Tianjin University, Tianjin, China
[6]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China
[7]Faculty of Surveying and Information Engineering, West Yunnan University of Applied Sciences, Yunnan, China

*Corresponding Author: Guangsheng Xie. Email: guangshengx@wyuas.edu.cn

Received: 29 October 2025; Accepted: 13 February 2026; Published: 15 June 2026

ABSTRACT: The attention mechanism, as a key technology for enhancing the performance of deep learning, is gaining increasingly widespread attention in medical image analysis due to its ability to focus on critical features and suppress redundant information. In recent years, the continuous evolution of attention methods has significantly improved their accuracy and robustness in key medical tasks such as lesion detection, tissue segmentation, and multimodal fusion, providing crucial support for building reliable clinical decision support systems. This paper systematically reviews the advances in attention-based methods for medical image analysis, comparing their performance with mainstream models like U-shaped convolutional neural network (UNet), convolutional neural network (CNN), and Vision Transformer (ViT) across multiple tasks, and summarizes various evaluation metrics, including the Dice coefficient, area under the ROC (Receiver Operating Characteristic) curve (AUC), and mean Average Precision (mAP). The review highlights that attention mechanisms bring significant improvements across core tasks such as segmentation, classification, detection, registration, and multimodal fusion: the Dice coefficient increases by 5%–12% in segmentation tasks, AUC improves by 3%–8% in classification tasks, mAP rises by 7%–15% in detection tasks, alignment accuracy enhances by 10%–20% in registration tasks, and retrieval accuracy reaches 85%–95% in multimodal fusion. The design characteristics and performance gains of key architectures such as channel attention, spatial attention, and hybrid attention are further analyzed. Nevertheless, current research still faces critical challenges, including scarce annotated data, limited cross-center generalization, complexity in multimodal fusion, insufficient model interpretability, and high computational costs. Accordingly, future research directions are proposed to promote the in-depth development and clinical translation of attention mechanisms in medical image analysis.

KEYWORDS: Attention mechanism; medical imaging; deep learning; multimodal fusion; explainability

1 Introduction

1.1 Background and Limitations of Medical Image Analysis

Medical imaging plays a central role in contemporary healthcare, supporting disease screening, diagnosis, treatment planning, intraoperative navigation, and follow-up assessment [1,2]. With advances in

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

modalities such as MRI, CT, PET, ultrasound, endoscopy, and high-resolution microscopy, the amount of image data available for clinical and research use has increased dramatically [3]. However, medical images are typically high-resolution and exhibit complex anatomy, small and sparsely distributed lesions, modality heterogeneity, organ variability, and various noise sources, which makes reliable extraction of lesion and structural information difficult. These challenges arise from the inherent differences between medical images and natural images. Medical images are predominantly grayscale, exhibit low contrast between different tissues, and often contain small and sparse targets (such as lesions or microcalcifications) embedded within complex anatomical backgrounds. Moreover, medical imaging frequently involves volumetric data, which requires modeling long-range dependencies across slices or temporal points. These characteristics impose stringent requirements on analysis methods, demanding precise localization of subtle abnormalities, preservation of anatomical topological structures, and efficient processing of high-resolution volumetric data. Traditional medical image analysis approaches—including handcrafted feature extraction, classical machine learning models, and early segmentation frameworks—have made progress in fundamental tasks but suffer from limited representational capacity and poor generalization ability. Furthermore, their heavy reliance on expert-designed features restricts their effectiveness in handling multimodal data, tiny lesions, low-contrast cases, and large-scale datasets. These limitations have driven the development of more expressive and adaptive intelligent analysis methods, particularly attention mechanisms that can highlight clinically relevant regions, suppress irrelevant background information, and model global context in a resource-efficient manner. Traditional medical image analysis methods, including hand-crafted features and classical machine-learning models [4], as well as early segmentation frameworks [5–8], have achieved progress in basic tasks but suffer from limited representation capacity and poor generalization. However, their strong reliance on expert-designed features limits their ability to cope with multimodal data, tiny lesions, low-contrast cases, and large-scale datasets, thereby motivating the development of more expressive and adaptive intelligent analysis methods.

In recent years, deep learning, particularly convolutional neural networks, has significantly advanced progress in medical image classification [9], segmentation [10], detection [11], registration [12], and generation. Nevertheless, performance and clinical applicability are still limited by data, models, and deployment constraints. These challenges can be broadly divided into three aspects. (a) Data Level: Medical datasets often suffer from scarce and expensive annotations, cross-centre heterogeneity, and class imbalance, which together hinder stable training and cause noticeable performance degradation when the data distribution shifts. (b) Model Level: Conventional deep networks still struggle to highlight tiny lesions against complex backgrounds, to effectively fuse heterogeneous multimodal information, and to provide transparent decision processes. Their performance is sensitive to small training sets, and processing high-resolution images incurs substantial computational and memory costs. (c) Clinical Level: Real-time deployment is constrained by hardware resources and latency, while a lack of interpretability, prospective validation, and robustness across devices and institutions remains a major obstacle to routine clinical adoption. These limitations underline the need for new architectures that can more flexibly focus on informative regions, model long-range dependencies, and adapt to large, heterogeneous datasets.

1.2 The Rise and Potential of Attention Mechanisms

To address these challenges, attention mechanisms are emerging as a key research direction in the field of medical image analysis. Inspired by the human visual system's ability to selectively focus on critical information within complex scenes, they simulate the automatic process of human visual perception that concentrates on important areas. This concept was first widely applied in the field of natural language processing, such as in the attention mechanisms within Seq2Seq models, and was revolutionized by the

Transformer architecture [13]. The subsequent introduction of the Vision Transformer (ViT) [14] marked a breakthrough in computer vision, as it divided images into patch sequences for global context modeling, freeing models from the constraints of convolutional priors. Swin Transformer further advanced this by introducing hierarchical structures and sliding window attention, effectively balancing computational efficiency with local and global feature modeling [15]. This progression made Transformers particularly suitable for high-resolution, dense prediction tasks like medical imaging. Building on these advancements, researchers have adapted various structural modifications and task-specific adaptations for medical imaging scenarios. For instance, TransUNet combines the local texture capture capability of convolutional neural networks with the global dependency modeling of Transformers to achieve more precise medical image segmentation [16]. Swin-Unet leverages window self-attention and hierarchical feature fusion mechanisms to significantly enhance the model's performance in multi-scale feature representation and cross-modal fusion [17]. These models effectively improve medical image analysis in terms of feature representation, structural recognition, and model interpretability. Attention mechanisms demonstrate formidable application potential through their exceptional feature selection and information aggregation capabilities. Their main advantages are as follows: (1) Automatically highlights lesion regions while suppressing background noise, enhancing detection capabilities for minute abnormal structures [18]. (2) Cross-attention and correlation variations enable alignment of heterogeneous modalities, facilitating accurate multimodal diagnosis [19]. (3) Attention maps provide intuitive visual cues for model decision-making, enhancing transparency and clinician trust [20]. (4) More efficient feature representations mitigate the impact of limited annotations and distribution shifts, improving transferability and generalization. (5) Sparse attention significantly reduces computational costs for high-resolution images while maintaining performance [21].

Leveraging these advantages, numerous attention-based model variants have been widely applied to various critical medical tasks. For instance, in pathological image classification, attention mechanisms effectively focus on diagnostically valuable regions within tissue sections and thereby enhance classification accuracy and robustness [22]. In tumor segmentation, multi-frequency multi-scale attention networks and generalized segmentation methods substantially enhance models' recognition and generalization capabilities for complex tissue structures [23,24]. In the task of detecting pulmonary nodules, the synergistic interaction between local discriminative features and attention mechanisms enables sensitive detection of small lesions [11]. In preoperative and intraoperative image registration, attention-guided unsupervised multimodal registration methods improve spatial alignment accuracy across different modalities and time points, providing technical assurance for surgical planning and efficacy evaluation [25]. Furthermore, in medical image automatic report generation tasks, cross-modal contrastive attention and vision-language pre-trained models effectively promote deep integration between clinical text and image features, enhancing report generation accuracy and interpretability [26,27]. Overall, attention mechanisms have significantly enhanced model accuracy and interpretability through their exceptional capabilities in information aggregation, feature selection, and context modeling. They have also strengthened the adaptability and generalization capabilities of models in practical clinical scenarios. Consequently, attention mechanisms have emerged as a core technological direction driving the advancement of intelligent medical image analysis and provide a new paradigm for intelligent medicine that bridges algorithmic innovation with clinical translation.

1.3 Motivation and Contributions of This Review

Despite significant advances in attention mechanisms for medical image analysis, numerous technical bottlenecks persist in practical clinical applications. These challenges encompass difficulties in multimodal fusion, interpretability, cross-domain generalization, and computational efficiency. While numerous reviews have explored deep learning applications in medical image analysis, they often lack a unified framework to

systematize the rapidly evolving attention mechanisms. They fail to conduct in-depth analyses of the trade-offs between performance and complexity, and offer limited insights into their integration with emerging large-scale foundational models.

To address this challenge, this study systematically summarizes the fundamental principles of attention mechanisms and proposes a novel classification framework. Representative studies across major medical image analysis tasks are reviewed, current technical bottlenecks are analyzed, and future development trends are projected to provide researchers with structured reference and guidance. Unlike discussions that often focus on a single perspective, this review emphasizes the technical principles, model design, downstream task applications, and clinical interpretability of attention mechanisms, thereby providing systematic guidance for further advancements in the field. The specific contributions of this review are reflected in the following four aspects:

(1) A unified multi-perspective classification framework for attention mechanisms. A novel structured taxonomy is proposed to systematically categorize attention mechanisms across multiple orthogonal dimensions: (a) distinguishing soft attention from hard attention structurally; (b) dimension-based classification into channel attention, spatial attention, spatio-temporal attention, and branch attention; (c) scope-based distinction between local attention and global attention; (d) architectural description of hybrid and hierarchical attention mechanisms. This framework offers a comprehensive perspective for understanding and comparing different attention paradigms.

(2) Comprehensive performance benchmarking and critical analysis. An extensive review of attention-based models' performance across core medical image analysis tasks is provided, including segmentation, classification, detection, registration, image generation, and multimodal fusion. A key contribution of this review is a critical comparison of these models against traditional convolutional neural networks and Transformer models, summarizing quantifiable performance gains while explicitly discussing computational costs, hardware requirements, and practical deployment limitations reported in recent literature.

(3) Comprehensive coverage of emerging and advanced applications of attention mechanisms. Beyond traditional classification and segmentation tasks, cutting-edge applications in unsupervised image registration, image generation and augmentation, and multimodal vision-language understanding are systematically investigated. This highlights the pivotal role of advanced mechanisms like cross-attention in achieving semantic alignment and fusion across heterogeneous data sources.

(4) Critical outlook on challenges and integration with foundational models. A forward-looking analysis is provided, delving into persistent challenges such as data scarcity, domain transfer, and computational burden. Crucially, specific future research directions are proposed, with particular discussion of integrating attention mechanisms into large-scale pre-trained models, thereby charting a course toward building unified and trustworthy clinical AI systems.

Subsequent sections will elaborate on the latest advancements and technical challenges in this domain across the following dimensions: Foundations of Attention Mechanism (Section 2), Application of Attention Mechanisms in Medical Image Analysis (Section 3), Challenges and Future Research Directions (Section 4). Section 5 presents the concluding summary of the entire paper.

2 Foundations of the Attention Mechanism

The attention mechanism was first introduced in Neural Turing Machines as a content-based addressing scheme [28]. It was soon adapted to sequence-to-sequence models to alleviate the information bottleneck in the standard encoder-decoder architecture [29,30]. In these models, compressing an entire input sequence into a single fixed-length vector often causes earlier information to be forgotten. Attention overcomes this limitation by allowing the decoder to dynamically attend to all encoder hidden states when generating each output token.

In Fig. 1, the input sequence $\{x^{(1)}, \ldots, x^{(T)}\}$ is mapped to encoder hidden states $\{h^{(1)}, \ldots, h^{(T)}\}$, which form the memory of the model. At each decoding step, the attention mechanism computes a context vector c that summarizes relevant encoder states. The decoder then updates its hidden states $\{h'^{(1)}, \ldots, h'^{(T')}\}$ and produces outputs $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(T')}\}$ conditioned on both the previous tokens and the context vector, enabling flexible and context-aware sequence generation.

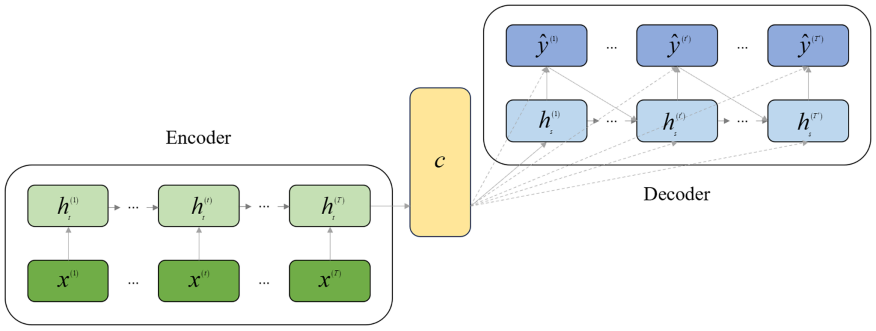


Figure 1: Encoder-decoder architecture.

2.1 Principles

Attention mechanisms, inspired by human perception, allow deep models to focus on salient parts of the input and suppress irrelevant information. A general formulation is

$$\text{Attention} = f(g(\mathbf{x}), \mathbf{x}), \tag{1}$$

where $g(\mathbf{x})$ generates an attention map or weighting over the input, and $f(g(\mathbf{x}), \mathbf{x})$ combines this map with $\mathbf{x}$ to emphasize important regions [31].

This view unifies many attention variants. For example, self-attention [13] and squeeze-and-excitation (SE) attention can be written as:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Linear}(\mathbf{x}) \tag{2}$$

$$g(\mathbf{x}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right) \tag{3}$$

$$f(g(\mathbf{x}), \mathbf{x}) = g(\mathbf{x})\mathbf{V} \tag{4}$$

for self-attention, and

$$g(\mathbf{x}) = \text{Sigmoid}(\text{MLP}(\text{GAP}(\mathbf{x}))) \tag{5}$$

$$f(g(\mathbf{x}), \mathbf{x}) = g(\mathbf{x}) \odot \mathbf{x} \tag{6}$$

for SE attention, where $\odot$ denotes element-wise multiplication with broadcasting, $\mathrm{GAP}(\mathbf{x})$ denotes global average pooling over spatial dimensions to produce a channel descriptor, and $\mathrm{MLP}(\cdot)$ denotes a multilayer perceptron (MLP) (a small fully-connected network) that maps this descriptor to channel-wise gating weights.

The seminal work of Vaswani et al. [13] further formalized attention as a mechanism consisting of three steps: (1) measure compatibility between a query and keys, (2) normalize these scores into attention weights, (3) compute a weighted sum of the corresponding values.

The most widely used instantiation is scaled dot-product attention. Given queries $\mathbf{Q} \in \mathbb{R}^{n_q \times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n_k \times d_k}$, and values $\mathbf{V} \in \mathbb{R}^{n_k \times d_v}$, the output is

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{7}$$

where the factor $1/\sqrt{d_k}$ stabilizes gradients by avoiding overly large dot products.

Multi-head attention extends this by computing several attention maps in parallel. The detailed procedure of multi-head attention is summarized in Algorithm 1.

Algorithm 1: Multi-Head Attention.

Input: Input sequence $X \in \mathbb{R}^{n \times d_{\text{model}}}$, number of heads h, projection matrices W_Q, W_K, W_V, W_O.

Output: Output sequence $Y \in \mathbb{R}^{n \times d_{\text{model}}}$.

// Project to queries, keys, and values

$Q \leftarrow XW_Q; \quad K \leftarrow XW_K; \quad V \leftarrow XW_V;$

// Split into h heads

$Q^{(i)}, K^{(i)}, V^{(i)}$ for $i = 1, \ldots, h;$

for $i \leftarrow 1$ **to** h **do**

$S^{(i)} \leftarrow \frac{Q^{(i)}(K^{(i)})^\top}{\sqrt{d_k}};$

$A^{(i)} \leftarrow \mathrm{Softmax}(S^{(i)});$

$Z^{(i)} \leftarrow A^{(i)}V^{(i)};$

// Concatenate heads and project

$Z \leftarrow \mathrm{Concat}(Z^{(1)}, \ldots, Z^{(h)});$

$Y \leftarrow ZW_O;$

// Optional residual connection and layer norm

$Y \leftarrow \mathrm{LayerNorm}(X + Y);$

return $Y;$

The Transformer [13] fully exploits self-attention by replacing recurrence with stacked encoder and decoder blocks, each combining multi-head attention with a position-wise feed-forward network. The overall Transformer architecture is presented in Algorithm 2.

Algorithm 2: The Transformer Architecture.

Input: Source sequence $\mathbf{x} = (x_1, \ldots, x_n)$, Target sequence $\mathbf{y} = (y_1, \ldots, y_m)$
Output: Output probabilities for the next token
// Encoder
$\mathbf{H}^0_{\text{enc}} \leftarrow \text{Embedding}(\mathbf{x}) + \text{PositionalEncoding}(\mathbf{x})$;
for $l \leftarrow 1$ **to** N **do**
  $\mathbf{H}^l_{\text{enc}} \leftarrow \text{EncoderLayer}(\mathbf{H}^{l-1}_{\text{enc}})$;
$\mathbf{C} \leftarrow \mathbf{H}^N_{\text{enc}}$;
// Decoder
$\mathbf{H}^0_{\text{dec}} \leftarrow \text{Embedding}(\mathbf{y}) + \text{PositionalEncoding}(\mathbf{y})$;
for $l \leftarrow 1$ **to** N **do**
  $\mathbf{H}^l_{\text{dec}} \leftarrow \text{DecoderLayer}(\mathbf{H}^{l-1}_{\text{dec}}, \mathbf{C})$;
// Final Output Layer
$\text{Logits} \leftarrow \text{Linear}(\mathbf{H}^N_{\text{dec}})$;
$\text{Probabilities} \leftarrow \text{Softmax}(\text{Logits})$;
return *Probabilities*;

The Transformer encoder layer applies multi-head self-attention followed by a feed-forward network, both wrapped with residual connections and layer normalization. The complete computation of the encoder layer is described in Algorithm 3.

The decoder layer extends the encoder design by incorporating three components: masked self-attention to prevent information leakage from future tokens, cross-attention to integrate encoder outputs, and a feed-forward network. Algorithm 4 summarizes the decoding process.

Algorithm 3: Encoder layer

Input: Input from previous layer $\mathbf{Z} \in \mathbb{R}^{n \times d_{\text{model}}}$
Output: Output of the encoder layer $\in \mathbb{R}^{n \times d_{\text{model}}}$
// Multi-Head Self-Attention
$\mathbf{A} \leftarrow \text{MultiHeadAttention}(\text{query} = \mathbf{Z}, \text{key} = \mathbf{Z}, \text{value} = \mathbf{Z})$;
$\mathbf{Z}' \leftarrow \text{LayerNorm}(\mathbf{Z} + \mathbf{A})$;
// Position-wise Feed-Forward Network
$\mathbf{F} \leftarrow \text{FeedForward}(\mathbf{Z}')$;
$\text{Output} \leftarrow \text{LayerNorm}(\mathbf{Z}' + \mathbf{F})$;
return *Output*;

Algorithm 4: Decoder layer

Input: Input from previous decoder layer $\mathbf{Z} \in \mathbb{R}^{m \times d_{\text{model}}}$, Encoder output $\mathbf{C} \in \mathbb{R}^{n \times d_{\text{model}}}$
Output: Output of the decoder layer $\in \mathbb{R}^{m \times d_{\text{model}}}$
// Masked Multi-Head Self-Attention
$\mathbf{A}_{\text{self}} \leftarrow \text{MultiHeadAttention}(\text{query} = \mathbf{Z}, \text{key} = \mathbf{Z}, \text{value} = \mathbf{Z}, \text{mask} = \text{future})$;
$\mathbf{Z}' \leftarrow \text{LayerNorm}(\mathbf{Z} + \mathbf{A}_{\text{self}})$;
// Multi-Head Cross-Attention
$\mathbf{A}_{\text{cross}} \leftarrow \text{MultiHeadAttention}(\text{query} = \mathbf{Z}', \text{key} = \mathbf{C}, \text{value} = \mathbf{C})$;

(Continued)

Algorithm 4 (continued)

$\mathbf{Z}'' \leftarrow \text{LayerNorm}(\mathbf{Z}' + \mathbf{A}_{\text{cross}})$;
// `Position-wise Feed-Forward Network`
$\mathbf{F} \leftarrow \text{FeedForward}(\mathbf{Z}'')$;
Output $\leftarrow \text{LayerNorm}(\mathbf{Z}'' + \mathbf{F})$;
return *Output*;

In sequence-to-sequence models, the decoder hidden state $\mathbf{h}_t$ serves as the query, and encoder hidden states $\{\bar{\mathbf{h}}_s\}$ as keys and values. The alignment weights are computed as

$$a_t(s) = \frac{\exp(\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s))}{\sum_{s'} \exp(\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_{s'}))}, \tag{8}$$

$$\mathbf{c}_t = \sum_s a_t(s)\bar{\mathbf{h}}_s, \tag{9}$$

where $a_t(s)$ measures the relevance of source position s at decoding step t, and $\mathbf{c}_t$ is the resulting context vector.

Common scoring functions include

$$\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s) = \begin{cases} \mathbf{h}_t^\top \bar{\mathbf{h}}_s & \text{(dot)} \\ \mathbf{h}_t^\top \mathbf{W}_a \bar{\mathbf{h}}_s & \text{(general)} \\ \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\mathbf{h}_t; \bar{\mathbf{h}}_s]) & \text{(concat)} \end{cases} \tag{10}$$

and the context vector is typically combined with $\mathbf{h}_t$ as $\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_c[\mathbf{c}_t; \mathbf{h}_t])$ for prediction [13].

Within the unified formulation, attention mechanisms share several key properties. In self-attention, $g(\mathbf{x})$ produces content-dependent weights across all positions and $f(g(\mathbf{x}), \mathbf{x})$ aggregates them, giving a global receptive field and strong long-range dependency modeling [13,31]. Stand-Alone Self-Attention showed that replacing spatial convolutions with such global attention can yield better accuracy–floating-point operations (FLOPs, a common proxy for computational cost) trade-offs on ImageNet (a large-scale image classification benchmark) and COCO (a standard benchmark for object detection/segmentation), especially in deeper layers [32]. ViT further demonstrated that pure Transformers without convolutions can match or surpass convolutional neural networks (CNNs) on large-scale image classification, underscoring the scalability of global modeling in data-rich regimes [14].

Attention matrices also provide useful, though not sufficient, tools for interpretation. Analyses of Bidirectional Encoder Representations from Transformers (BERT, a bidirectional Transformer-based language model)'s attention heads revealed stable patterns linked to positional alignment, syntax, and coreference, suggesting meaningful internal structure [33]. Yet attention weights alone may not reflect causal importance. Layer-wise rollout methods that combine attention with relevance propagation and skip connections yield more stable and task-aligned explanations than raw attention visualization, indicating that robust explanation typically requires combining attention with propagation-based saliency [34].

Finally, attention and convolution exhibit complementary strengths: convolution supplies locality, translation equivariance, and strong inductive bias, whereas self-attention provides content-driven global modeling with adaptive receptive fields. Empirically, self-attention can either replace high-level convolutions [32] or form pure Transformer backbones such as ViT [14]. In practice, this leads to three main design patterns: adding attention modules on top of CNN backbones (enhancement), replacing large-receptive-field

convolutions in higher layers with attention (hybrid), or adopting fully Transformer-based architectures, each chosen according to task requirements and computational constraints.

2.2 Taxonomy of Attention Mechanisms

In deep learning, attention mechanisms can be organized into a structured taxonomy that clarifies their design principles and application scopes. As shown in Fig. 2, they can be grouped along four dimensions: operational dimension, structure, receptive-field scope, and hybrid/hierarchical combinations. Dimension-wise attention covers channel, spatial, spatio-temporal, and branch attention. Structure-wise attention distinguishes soft (continuous) and hard (discrete) weighting. Scope-wise attention separates local from global receptive fields. Recent advances further introduce CNN-Transformer fusion, hierarchical attention, and cross-layer feature aggregation to enhance representation capacity and flexibility.

Rather than viewing the taxonomy as a list of modules, it is intended as a decision aid: (1) choose the operational dimension to match task needs, (2) choose soft/hard structure to match supervision, and (3) choose local/global scope to match resolution and compute budget.

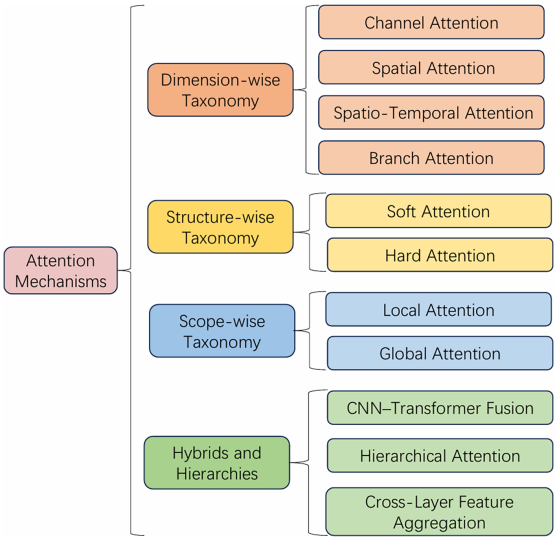


Figure 2: Taxonomy of attention mechanisms.

2.2.1 Dimension-Wise Taxonomy

This subsection introduces a dimension-wise taxonomy that categorizes attention mechanisms according to the primary feature dimension they operate on. Four major categories are considered: channel attention, spatial attention, spatio-temporal attention, and branch attention. Fig. 3 illustrates the workflows of the first three: (a) channel attention models inter-channel dependencies via global pooling and shared MLPs, (b) spatial attention emphasizes important regions using pooled feature maps and convolutional filtering, and (c) spatio-temporal attention combines channel and spatial attention to capture joint variations across space and time. Table 1 further summarizes their main characteristics, motivations, and typical use cases.

Beyond definitions, dimension-wise attention can be understood as a set of design trade-offs: channel attention mainly re-weights *what* to emphasize (semantic/feature selectivity) with low overhead, whereas spatial attention highlights *where* to focus (localization) but is more sensitive to resolution and receptive-field design. In medical imaging, these trade-offs are further shaped by small-lesion sparsity, 3D volume resolution, and domain shift, which motivates task-driven alignment principles.

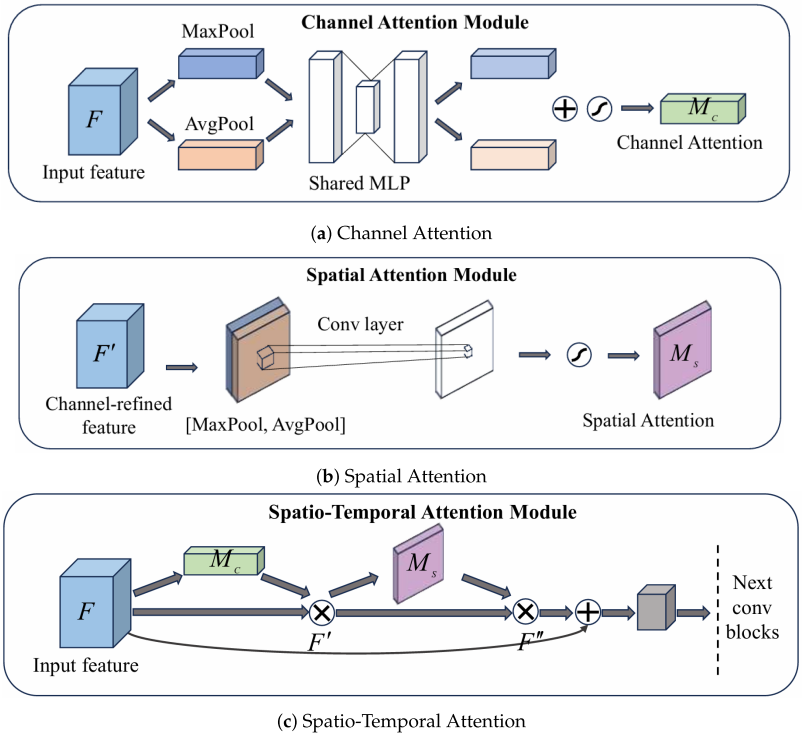


Figure 3: Illustration of dimension-wise attention mechanisms: (**a**) channel, (**b**) spatial, and (**c**) spatio-temporal attention.

Table 1: A comparative summary of attention mechanisms based on their primary operational dimension.

Attention Type	Advantages	Limitations	Typical Applications
Channel Attention	Models interdependencies between feature channels. Lightweight, easy to integrate. Favorable accuracy-parameter trade-off.	Lacks spatial modeling capability. May suppress informative channels if not regularized.	Image classification (ImageNet, CIFAR). Fine-grained visual categorization (Caltech–UCSD Birds (CUB)). Medical image analysis (ISIC, BraTS).
Spatial Attention	Focuses on semantically informative regions. Enhances localization for dense prediction tasks. Can capture multi-scale contextual information.	Computational cost increases quadratically with spatial resolution. Ignores channel-wise relationships. Performance can be sensitive to kernel/window size.	Object detection (MS COCO). Semantic segmentation (Cityscapes, ADE20K).
Spatio-Temporal Attention	Models complex dependencies across both space and time. Essential for understanding dynamic visual content. Architecture can be factorized for efficiency.	Vanilla self-attention has quadratic complexity in space-time. High memory consumption for long sequences. Can be prone to training instability.	Video action recognition (Kinetics, UCF101). Time-series forecasting. Autonomous driving trajectory prediction.

(Continued)

Table 1 (continued)

Attention Type	Advantages	Limitations	Typical Applications
Branch Attention	Dynamically selects network branches or features. Adapts model capacity to input scale/complexity. Can improve robustness and representation power.	Introduces additional parameters and inference latency. Risk of branch collapse without proper regularization. Requires careful hyperparameter tuning.	Multi-scale feature fusion (SKNet). Architecture search, Mixture-of-Experts. Tasks requiring dynamic inference.

(1) Channel Attention

The channel attention mechanisms model interdependencies between feature channels and adaptively rescale them, amplifying informative channels and suppressing redundant ones. A typical 'squeeze-and-excitation" pipeline consists of: (i) *Squeeze*: global spatial information is aggregated into a channel descriptor using global average pooling (GAP); (ii) *Excitation*: a small gating network produces a weight for each channel; (iii) *Reweighting*: the input features are multiplied by these weights. Let $X \in \mathbb{R}^{C\times H\times W}$ be an input feature tensor. The squeeze step produces $\mathbf{z} \in \mathbb{R}^{C}$, where $z_c = \mathrm{GAP}(X_c)$. In the SE block [35], excitation and reweighting are

$$\mathbf{a} = \sigma(\mathbf{W}_2\,\delta(\mathbf{W}_1\mathbf{z})), \quad Y = \mathbf{a}\odot X, \tag{11}$$

with rectified linear unit (ReLU) δ, Sigmoid σ, parameters $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, reduction ratio r, and channel-wise multiplication $\odot$. Representative variants include:

(i) Squeeze-and-Excitation (SE) [35]: the original squeeze-excitation block with fully-connected layers.
(ii) Efficient Channel Attention (ECA) [36]: replaces fully connected (FC) layers with 1D convolution for local cross-channel interaction.
(iii) Convolutional Block Attention Module (CBAM) [37]: cascades channel and spatial attention.
(iv) Selective Kernel (SKNet) [38]: uses channel-wise attention to select among convolution kernels of different sizes.

Channel attention is powerful and efficient, but global average pooling discards spatial information by collapsing each channel into a single scalar, so the model learns 'what" is important but not precisely where" it occurs, which is limiting for dense prediction [35,37]. Moreover, the fully-connected gating in SE can introduce nontrivial parameter overhead. Future work, therefore, focuses on richer aggregation that preserves partial spatial context, lighter cross-channel interaction as in ECA-Net [36], and designs where channel weights are conditioned on local spatial cues, thereby narrowing the gap between channel and spatial attention.

(2) Spatial Attention

Spatial attention works on the spatial dimensions and aims to highlight informative regions in the feature map. It generates a 2D mask $M_s \in \mathbb{R}^{H\times W}$ shared across channels:

$$M_s = \sigma\left(f^{k\times k}\left([\mathrm{AvgPool}_c(X);\mathrm{MaxPool}_c(X)]\right)\right), \quad Y = M_s\odot X, \tag{12}$$

where σ is sigmoid, $f^{k\times k}$ is a $k\times k$ convolution, and $\mathrm{AvgPool}_c$, $\mathrm{MaxPool}_c$ denote pooling along channels. More advanced variants use multi-scale or dilated convolutions, deformable sampling, or Transformer-style self-attention over flattened patches.

Spatial attention is widely adopted in medical image analysis to model spatial transformations and long-range contextual dependencies. VoxelMorph [39] introduces a learning-based framework for deformable medical image registration, enabling differentiable spatial transformations that adapt to complex anatomical variability. To efficiently capture long-range spatial context under computational constraints, Medical Transformer (MedT) [40] employs gated axial attention, decomposing full self-attention into separable spatial dimensions while preserving global dependency modeling.

The main challenge of spatial attention is balancing receptive field and computational cost. Convolution-based designs are limited by local receptive fields, whereas Transformer-style attention captures global context but incurs quadratic complexity in spatial resolution, which is costly for high-resolution medical images [41]. Moreover, a single spatial mask shared across channels may be overly restrictive [42]. Promising directions include efficient global modeling (e.g., linear or sparse attention), channel-aware spatial modulation, and deformable or adaptive sampling to better fit irregular anatomical structures.

(3) Spatio-Temporal Attention

Spatio-temporal attention extends attention to sequential data such as videos and time series by jointly modeling dependencies across space and time. To control complexity, many architectures factorize or approximate full attention, for example, by applying separate spatial and temporal attention, using local temporal windows with global spatial attention, or introducing memory tokens and streaming caches:

$$Y = \mathrm{Attn}_t\left(\mathrm{Attn}_s(X)\right) \quad \text{or} \quad \mathrm{Attn}_{st}(Q, K, V) \tag{13}$$

with $(H \times W \times T)$ tokenization.

Non-local Neural Networks were early adopters of global spatio-temporal attention for video understanding. Subsequent work applied similar ideas to skeleton-based action recognition, trajectory prediction, action recognition and detection, and multivariate time-series forecasting.

Spatio-temporal attention is crucial for dynamic data but remains computationally heavy because full self-attention scales quadratically with the number of space–time tokens and often requires large training datasets [43]. Research is therefore moving toward linear or sparse attention, hierarchical designs that capture local and global relations at different scales, and adaptive temporal schemes that focus computation on key frames or change points, improving both efficiency and data utilization.

(4) Branch Attention

Branch attention dynamically selects or combines features from parallel branches (e.g., convolutional kernels or experts). Given m branch transformations $\{B_j(X)\}_{j=1}^{m}$, a gating network produces weights $\mathbf{g} \in \mathbb{R}^m$:

$$\mathbf{g} = \mathrm{softmax}\left(\mathbf{U}\,\phi(\mathrm{Pool}(X))\right), \quad Y = \sum_{j=1}^{m} g_j \cdot B_j(X), \tag{14}$$

where $\mathbf{U}$ is a projection matrix, ϕ a nonlinearity, and Pool a global pooling function. This formulation underlies SKNet, Mixture-of-Experts (MoE) models, and dynamic convolution.

The Attention Branch Network (ABN) [44] introduces a dedicated attention branch that generates spatial attention maps to modulate the perception branch, enabling both improved classification performance and enhanced interpretability in medical image analysis, as demonstrated in panoramic radiograph–based dental implant classification. Branch attention has also been extended to multi-branch medical image segmentation, such as dual-branch Transformer–CNN architectures with dual attention mechanisms that explicitly model global contextual information and local fine-grained features and fuse them through

attention-guided branch interactions [45], as well as 3D multi-branch attention for brain tumor segmentation (MBANet) [46].

Branch attention improves flexibility but increases computational and memory cost because multiple branches and a gating network are evaluated [47]. Poorly designed gating can cause branch collapse or overfitting. Future work aims at lighter, more robust gating (e.g., hierarchical or coarse-to-fine strategies) and tighter integration with other attention types, such as using dedicated branches for cross-modal fusion or spatial reasoning, while keeping the overall design deployable in resource-constrained clinical environments.

2.2.2 Structure-Wise Taxonomy

This taxonomy distinguishes attention mechanisms by the nature of the weights: continuous (soft) vs. discrete (hard). Soft attention works well when a model needs smooth, end-to-end training, which is common in dense prediction with limited supervision. Hard attention is more appropriate when a discrete selection is desired to reduce computation or to encode explicit anatomical constraints. In medical imaging, hard decisions can be sensitive to domain shift and noise; a soft mask with a light gating step is often more robust.

(1) Soft Attention

Soft attention assigns differentiable weights to all input elements and is trainable with standard backpropagation. It underpins most modern attention-based models, including self-, cross-, and cross-modal attention. While softmax is the default normalization, alternatives such as sparsemax and entmax encourage sparsity, and temperature scaling can improve training stability.

Key milestones include the use of soft attention in neural machine translation [30], image captioning via "Show, Attend and Tell" [48], and the Transformer [13]. Residual Attention Networks [49] integrate attention with residual learning, and hierarchical attention networks exploit document structure.

Soft attention still suffers from quadratic complexity and possible attention dilution, where weights spread over irrelevant regions. Sparse variants alleviate this, but may be harder to train. Future work focuses on more efficient approximations (e.g., linear attention) and structurally sparse yet stable mechanisms, ideally guided by domain priors to produce sharper, clinically meaningful attention maps.

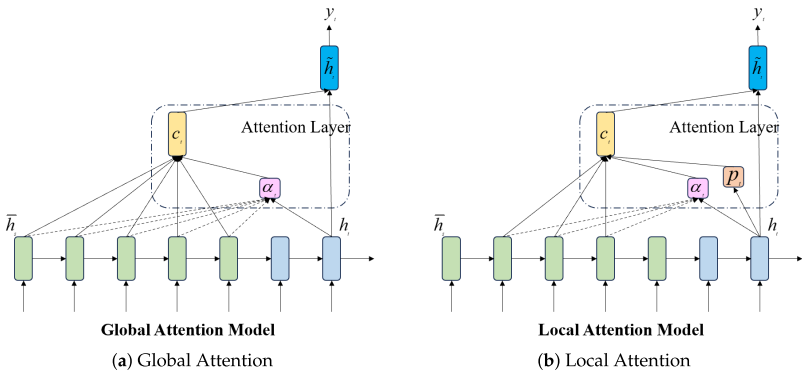
(2) Hard Attention

Hard attention makes discrete selections of patches, tokens, or branches, potentially reducing latency and FLOPs through explicit pruning or routing. It has shown value in efficiency- and interpretability-oriented tasks, including continual learning, multi-step visual reasoning, retinal vessel segmentation, and navigation for intelligent agents.

The main difficulties of hard attention are unstable training and potentially brittle decision boundaries, since small input changes can cause abrupt shifts in the selected regions [48]. Promising directions include better gradient estimators, hybrid soft–hard schemes such as differentiable masking, and anatomically guided selection in medical imaging to improve robustness and clinical plausibility.

2.2.3 Scope-Wise Taxonomy

This taxonomy classifies attention mechanisms by receptive field, distinguishing global and local attention. As illustrated in Fig. 4, global attention allows each position to interact with all others, whereas local attention restricts interactions to a neighborhood, improving efficiency.



(**a**) Global Attention (**b**) Local Attention

Figure 4: Comparison of (**a**) global and (**b**) local attention mechanisms. Key variables: $\overline{h}$ (encoder hidden states), h_t (decoder hidden state), c_t (context vector), α (attention weights), p_t (attention window center in local attention), $\tilde{h}_t$ (attentional hidden state), and y_t (output).

In Fig. 4, $\overline{h}$ denotes encoder hidden states, h_t the decoder hidden state at time t, and α the attention weights. The attention layer aggregates encoder states into a context vector c_t, which is combined with h_t to form the attentional hidden state $\tilde{h}_t$ used to predict y_t. Local attention further introduces p_t to mark the center of the attention window and limit computation to nearby positions.

(1) Global Attention

Global attention lets each token attend to all others and is well-suited for modeling long-range dependencies. To mitigate its quadratic cost, efficient variants introduce structured sparsity, kernelization, low-rank projections, k-nearest neighbors (kNN) routing, which selects a token's top-k most similar tokens as its attention targets, or hybrid local–global designs.

The Transformer [13] popularized global self-attention. Structured self-attention and hybrid models such as Longformer further reduced cost by mixing local patterns with global tokens.

In high-resolution or volumetric medical data, naive global attention is often too expensive, and fixed sparse patterns or kernel approximations may either restrict information flow or degrade performance [50].

For high-resolution 3D volumes, global attention is most effective when applied after downsampling or within a hierarchy; otherwise, memory and latency often dominate deployment feasibility.

(2) Local Attention

Local attention restricts the receptive field to a window of size w, reducing complexity from $O(n^2)$ to $O(nw)$:

$$\text{LocalAttn}(Q_t, K, V) = \text{Attn}\big(Q_t, K_{\mathcal{N}(t)}, V_{\mathcal{N}(t)}\big), \tag{15}$$

where $\mathcal{N}(t)$ is the neighborhood of position t. To improve information flow, designs often use shifted windows, dilated patterns, or cross-window connections.

Local attention has enabled attention models for long sequences and high-resolution inputs. Luong et al. [51] first proposed local attention for machine translation, extended by [52] with relative positions. Longformer [53] popularized sliding-window attention for long documents, and local self-attention has also been applied to image modeling [54].

The main limitation is the loss of global context when the window is too small, while increasing w reduces efficiency [14,15]. Future work includes adaptive window sizes that depend on content, hierarchical

architectures combining local attention at multiple scales, and hybrid schemes that couple sparse global tokens with dense local windows, ideally guided by anatomical priors in medical imaging.

A useful practice is to set the window size based on the typical scale of lesions and anatomical structures, and then add sparse global tokens or cross-window links to recover long-range dependencies.

2.2.4 Hybrids and Hierarchies

Real-world applications often benefit from combining different attention paradigms. Hybrid and hierarchical architectures integrate complementary mechanisms to support multi-scale feature aggregation and cross-modal interaction.

(1) CNN–Transformer Fusion

CNNs and Transformers are complementary: CNNs provide efficient locality in early stages, whereas Transformers capture long-range dependencies. Common fusion patterns include convolutional stems followed by Transformer blocks, interleaved convolution and attention layers, CNN encoders paired with Transformer decoders or heads, and integrated mixers that inject convolution into Transformer feed-forward layers.

Typical examples are CvT [55], which embeds convolution into tokenization and projection, Bottleneck Transformers [56], which replace bottleneck convolutions with self-attention, and efficient hybrids such as Mobile-Former [57] and LeViT [58].

Despite their strong performance, hybrid CNN-Transformer designs are often heuristic and can be heavy in parameters and memory, which complicates deployment in resource-limited clinics [59]. Future work aims at more principled fusion frameworks, possibly via neural architecture search or dynamic fusion that adapts the balance between convolutional and attention processing to each input, while respecting practical constraints on speed and memory.

In practice, fusion designs that keep early convolutional stages and reserve attention for higher-level semantics tend to offer a better accuracy–latency balance in clinical pipelines.

(2) Hierarchical Attention

Hierarchical attention models multi-scale information by building feature pyramids with patch merging or downsampling, applying attention within each level, and fusing cross-scale information. Global tokens or Classification Token ([CLS]) tokens aggregate semantics, and positional encodings maintain spatial consistency.

The Hierarchical Attention Network (HAN) [60] pioneered word- and sentence-level attention for document classification. Related ideas appear in structured self-attention [61] and downstream tasks such as relation classification and natural language inference.

Fixed hierarchical structures may not align with the semantic granularity of specific tasks and increase complexity, especially for high-resolution 3D medical images [62]. Future directions include more flexible hierarchies that adapt scale and interaction patterns based on the input, and anatomically informed decompositions where different levels correspond to meaningful biological scales.

Hierarchical attention is particularly advantageous when both global anatomical context and local boundary details are needed, such as organ–lesion joint modeling in 3D segmentation.

(3) Cross-Layer Feature Aggregation

Cross-layer aggregation combines low-level spatial details with high-level semantics from different depths. Typical strategies are: (i) top-down pathways with lateral connections (e.g., FPN); (ii) cross-stage attention between non-adjacent feature maps; (iii) gated skip connections that adaptively weight layers.

Representative architectures, such as Non-local Neural Networks [63] and SENet [35]. Stable training often relies on proper scaling, normalization, positional encodings, and efficiency techniques such as mixed-precision training, flash attention, and gradient checkpointing.

Cross-layer aggregation can bridge the semantic gap between low-level and high-level features while controlling memory usage. Simple fusion operations may not fully exploit complementary information, especially in medical imaging, where precise localization and global context are both critical. Future work will likely explore more intelligent, attention-based routing across layers, guided by anatomical hierarchies and assisted by efficient design tools such as neural architecture search and knowledge distillation to keep models compact enough for clinical deployment.

2.3 Requirement–Mechanism Alignment

Mechanism selection depends on task target scale, available supervision, and computational budget.

2.3.1 Task-Driven Dimension Selection

The choice of the operational dimension of attention can be guided by the requirements of the medical imaging task. Spatial attention is suitable for precise localization, such as small lesion detection or boundary delineation, as in Attention U-Net [64]. For recognizing complex multi-class tissue structures, channel attention mechanisms like SE blocks enhance discriminative channels. In dynamic imaging (e.g., ultrasound, cardiac MRI), spatio-temporal attention is needed to capture motion and temporal evolution, as in Transformer-based TransUNet [13,65]. When fusing multiple imaging pathways or modalities, branch attention and non-local operations [63,66] provide adaptive feature fusion. Aligning the attention dimension with the clinical objective is therefore a basic design principle.

2.3.2 Structure Selection under Supervision Conditions

The supervision regime also influences the choice between soft and hard attention. In weakly supervised or few-shot settings, where pixel-level labels are scarce, *soft attention* is preferred because it is differentiable and trainable end-to-end. Attention-based multiple instance learning [67] and Attention Gated Networks [47] are typical examples that leverage image-level labels to learn salient regions. When strong priors or explicit anatomical knowledge are available, *hard attention* can enforce discrete, focused processing. Inspired by early image captioning work [48], coarse-to-fine segmentation frameworks [68] use region-guided hard attention to refine predictions. Thus, the supervision paradigm naturally guides the choice of attention structure.

2.3.3 Computation and Resolution Trade-Offs

Medical image analysis often balances high spatial resolution with limited computational resources. For large 3D volumes, the quadratic cost of global self-attention is typically infeasible. Models such as UNETR [69] therefore adopt local or windowed attention to make Transformer-based segmentation tractable. When global context is critical, hierarchical architectures like TransBTS [70] and TransUNet [65] progressively aggregate multi-scale context. Hierarchical and multi-scale Transformer designs further alleviate this issue by progressively aggregating contextual information. For instance, HiFormer integrates multi-scale Transformer and convolutional representations through hierarchical fusion, enabling effective global modeling with reduced computational overhead [71]. Alternatively, attention factorization provides

a complementary efficiency strategy. Medical Transformer (MedT) employs gated axial attention to decompose full self-attention into separable spatial dimensions, substantially reducing memory and computational costs while retaining the ability to model long-range dependencies [40].

2.3.4 Multi-Modal and Cross-Domain Modeling

The integration of multi-modal data and robustness to domain shift are central challenges for clinical deployment. Attention facilitates adaptive alignment and fusion. For multi-modal learning, frameworks learn joint representations from unpaired CT and magnetic resonance imaging (MRI) [72] or synthesize magnetic resonance (MR) sequences via modality-invariant latent spaces [73]. For cross-domain adaptation, adversarial training with attention helps align feature distributions across imaging modalities or scanner domains [74,75], improving generalization to unseen data. In both cases, attention serves as a flexible tool to prioritize and integrate relevant signals from heterogeneous sources.

2.3.5 Interpretability and Clinical Trustworthiness

For clinical use, models typically need to be interpretable and trustworthy. Attention maps produced by models such as Attention Gated Networks [47] highlight regions that drive predictions and can be inspected by clinicians. Post-hoc attribution methods, such as saliency maps and class activation maps, are widely used to localize salient regions that contribute to model predictions in medical imaging, as systematically reviewed in [76], while generative-model-based visualization [77] further localizes salient features in arbitrary networks. Beyond visualization, incorporating anatomical priors into the attention process itself anchors the model in clinical knowledge and can improve both reliability and plausibility of explanations, as emphasized in comprehensive surveys [78].

3 Application of Attention Mechanisms in Medical Image Analysis

Self-attention and Transformer architectures have become core techniques in modern medical image analysis. By modeling long-range dependencies, reweighting informative features, and integrating heterogeneous modalities, attention alleviates the locality limitations of convolutional networks and improves both robustness and interpretability. Unlike natural image tasks where attention often serves to highlight semantically salient objects (e.g., cars, animals), medical image analysis demands attention mechanisms that are sensitive to clinically meaningful regions—subtle lesions, anatomical boundaries, functional anomalies—often under weak supervision, high class imbalance, and stringent robustness requirements. The following sections detail how attention architectures are specifically adapted to meet these medical-specific challenges across segmentation, classification, detection, registration, generation, and multimodal fusion and vision–language analysis.

As illustrated in Fig. 5, in image segmentation, models such as Attention U-Net, CBAM U-Net, TransAttUnet [79], TransUNet, Swin-UNet, MedT, UNETR, HiFormer, and nnFormer employ attention gating and Transformer-based encoders to achieve precise boundary delineation. Representative micro-lesion–oriented architectures include CaraNet [80], SvANet [81], STS-Net [82], MCANet [83], and SMAFormer, which focus on highlighting subtle targets [84].

In disease classification, models including SENet [35], SCSE [85], HiFuse [86], MedViT [87], MedViT-V2 [88], SwinECAT [89], CTransCNN [90], and CDDFuse [91] enhance feature representation through channel and spatial attention as well as multimodal fusion. In object detection and localization, frameworks such as EL-DETR [92], PMA-DETR [93], GravityNet [94], and MedYOLO [95] improve the accuracy of small-lesion detection under complex anatomical backgrounds.

Furthermore, models such as DaVoxelMorph [96], TransMorph [97], and DiffuseReg [98] for registration, DAGAN [99], SwinGAN [100,101], SwinMR [102], UniMIE [103], and MedDiffusion [104] for image generation and enhancement, and TransMed [105], MATR [106], DM-FNet [107], MedCLIP [108], MGCA [109], and BioViL-T [110] for multimodal fusion and vision-language modeling demonstrate strong cross-modal alignment and semantic reconstruction capabilities.

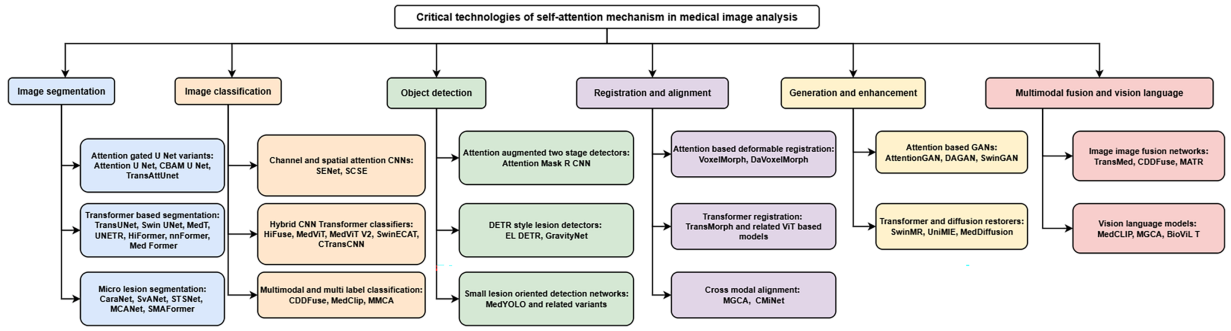


Figure 5: Critical technologies of self-attention mechanism in medical image analysis.

3.1 Medical Image Segmentation

Medical image segmentation aims to accurately delineate anatomical structures such as organs and tumors from complex images, providing a quantitative basis for surgical navigation, radiotherapy target delineation, and disease monitoring [111]. Among deep learning approaches, U-Net and its variants, remain the de facto backbone for many clinical segmentation pipelines. The original U-Net adopts an encoder–decoder architecture with skip connections that fuse high- and low-level features, and has demonstrated strong performance on a wide range of biomedical datasets [4]. However, the strictly local receptive field of convolution kernels limits the modeling of long-range dependencies and global context. To alleviate this limitation, attention mechanisms have been incorporated into the U-Net framework. Attention U-Net [64] introduces attention gates on the skip connections to suppress irrelevant background responses and to emphasize target regions, which improves the delineation of small structures and reduces false positives in cluttered backgrounds. Convolutional Block Attention Module (CBAM) U-Net [112] further combines channel and spatial attention modules with the U-Net to enhance feature representation and boundary localization, especially for small and low-contrast anatomical structures. In parallel, TransAttUnet [79] augments a U-Net-like encoder–decoder with Transformer blocks and multi-level attention, enabling better modeling of long-range dependencies while preserving local spatial detail. Transformer-based architectures have recently become a main research focus for medical image segmentation. For example, TransUNet [65] uses a hybrid CNN–Transformer encoder, in which a Vision Transformer (ViT) is stacked on top of a CNN feature extractor to encode global semantic context, and a U-shaped decoder recovers dense predictions. Swin-UNet [15] replaces the encoder with a hierarchical Swin Transformer that employs shifted-window self-attention, which substantially improves performance on the Synapse multi-organ CT dataset by capturing multi-scale context with reduced computational cost compared with naive global attention. Building on these ideas, HiFormer [113] employs a hierarchical dual-level Transformer, achieving leading Dice Similarity Coefficient (DSC) and significantly reducing HD_{95} among 2D methods on Synapse compared to TransUNet. Extending this to 3D, nnFormer [114] outperforms previous Transformer baselines by explicitly modeling volumetric context, yielding substantial improvements in both accuracy and boundary delineation, albeit with higher memory demands. Representative 2D attention- and Transformer-based segmentation models on the Synapse multi-organ CT dataset are summarized in Table 2.

Table 2: A comparative summary of representative models for medical image segmentation.

Model	DSC	HD_{95}	Hardware	Advantages	Limitations	Application	Ref.
U-Net	74.68	36.87	NVIDIA Titan (6 GB)	Classic encoder-decoder CNN; simple and easy to train; robust baseline.	Lacks explicit global context; strictly 2D slice-based; performance drops in complex scenarios.	General-purpose medical image segmentation.	[4]
Attn U-Net	75.57	36.97	Not reported	Attention gates on skip connections suppress noise and enhance small organ delineation.	Slightly increased parameters; still limited in long-range dependency modeling.	Precise localization of small abdominal organs.	[64]
TransUNet	77.48	31.69	RTX 2080Ti	Hybrid CNN-Transformer encoder; ViT bottleneck encodes global semantic context.	Higher memory usage and training time; limited full 3D context modeling.	Multi-organ segmentation prioritizing accuracy.	[65]
Swin-UNet	79.13	21.55	V100 (32 GB)	Pure Transformer; shifted-window attention captures multi-scale context.	High cost; sensitive to window size and input resolution.	High-accuracy segmentation on high-end Graphics Processing Units (GPUs).	[15]
HiFormer-B	80.39	14.70	RTX 3090	Hierarchical dual-level Transformer with cross-scale feature fusion.	Computationally heavy; less suitable for resource-limited deployments.	High-precision offline surgical planning.	[113]

Note: Dataset: Synapse multi-organ CT; Task: Multi-organ segmentation. Units: DSC in %; HD_{95} in mm.

All quantitative results in Table 2 are directly taken from the original publications and were not re-implemented under a unified experimental setup. The hardware configurations used for training also differ, as indicated in the "Hardware" column: for instance, the original U-Net was trained on an NVIDIA Titan GPU with 6 GB memory, TransUNet on a single NVIDIA RTX 2080Ti GPU, Swin-UNet on an NVIDIA V100 GPU with 32 GB memory, and HiFormer-B on an NVIDIA RTX 3090 GPU with 24 GB memory, whereas the Attention U-Net paper does not specify the GPU model. These differences in GPU type, memory capacity, and training settings affect both the reported DSC/HD_{95} values and the practical deployability of each model; Consequently, the numbers in Table 2 are best interpreted as indicative trends rather than strictly comparable benchmarks. From an application perspective, U-Net and Attention U-Net provide a favorable balance between accuracy and efficiency and can be trained and deployed on commodity GPUs, making them suitable for routine clinical workflows and near real-time use. Transformer-based hybrids such as TransUNet and Swin-UNet yield clear improvements in mean DSC and HD_{95} at the expense of substantially higher parameter counts, FLOPs, and memory usage, and are therefore better suited to offline

multi-organ delineation or research settings where high-end GPUs. High-end GPUs such as RTX 2080Ti and V100 are typically available in these settings, where latency is less critical. HiFormer-B further enhances segmentation accuracy on Synapse and achieves low HD_{95} among 2D slice-based methods, but its heavy computational footprint and reliance on GPUs such as RTX 3090 confine it to powerful workstations. When volumetric context is essential and sufficient GPU memory is available, fully 3D architectures such as nnFormer can further improve organ boundary consistency and small-structure delineation, but they exacerbate the trade-off between accuracy, computational complexity, and memory consumption.

3.2 Disease Classification and Recognition

3.2.1 Attention-Augmented Feature Extraction Networks

In medical image-based disease classification and recognition tasks, traditional convolutional neural networks (CNNs) excel at extracting local appearance features, but their inherently limited receptive fields restrict the modeling of long-range dependencies. As a result, purely convolutional backbones often struggle to capture the subtle, multi-scale manifestations of small lesions or diffuse disease patterns, especially on complex datasets such as dermoscopic images or chest radiographs. To tackle this issue, a series of attention mechanisms has been introduced to recalibrate feature responses and emphasize clinically relevant structures. SENet [35] employs a squeeze-and-excitation scheme to model the global channel dependencies and adaptively reweight feature maps, and can be seamlessly plugged into existing CNN architectures as a lightweight, generic feature optimizer for medical imaging pipelines. Building on this idea, Spatial and Channel Squeeze & Excitation (SCSE) [85] extends squeeze-and-excitation to both channel and spatial dimensions by introducing channel Squeeze-and-Excitation (cSE), spatial Squeeze-and-Excitation (sSE), and spatial and channel Squeeze-and-Excitation (scSE) blocks. When integrated into fully convolutional networks, SCSE substantially improves the saliency of lesion regions and yields higher Dice scores for brain and whole-body organ segmentation on MALC [115] and Visceral datasets. These classic attention modules incur only a modest parameter overhead, making them attractive for resource-constrained clinical deployments, but they still rely on local convolutional context and cannot explicitly encode long-range semantic relations.

With the advent of vision Transformers, modeling of global context has been significantly enhanced. Recent works increasingly adopt a hybrid CNN–Transformer architectures to balance local inductive biases and global self-attention. HiFuse [86] designs a three-branch hierarchical multi-scale fusion structure, where the Hierarchical Feature Fusion (HFF) block adaptively aggregates local and global streams and the inverted residual MLP (IRMLP) compensates for the lack of a feed-forward network while maintaining stable gradients. On ISIC2018, the HiFuse-Base [86] model reaches an accuracy of 85.85% and an Area Under the Curve(AUC) of 95.85%, with approximately 82.49M parameters and 8.13 Giga Floating-Point Operations (GFLOPs), illustrating a favorable trade-off between performance and computational cost for skin-lesion classification. MedViT [87] further adopts a CNN–Transformer hybrid framework, where the Efficient Convolution Block (ECB) strengthens local perception, the Long-Term Block (LTB) captures global dependencies, and the Progressive Mixup Consistency (PMC) module improves adversarial robustness. MedViT-S [87] demonstrates strong generalization across the MedMNIST-2D [88] suite. Building on this, MedViTV2 [88] integrates Kolmogorov–Arnold Network (KAN) layers and Dilated Neighborhood Attention, improving accuracy on both standard and corrupted datasets while reducing computational FLOPs by approximately 44 percentage compared to its predecessor. MedViTV2 achieves state-of-the-art performance on 17 medical image classification datasets and 12 corrupted datasets, improving the average accuracy by up to 4.6% on MedMNIST and 13.4% on MedMNIST-C while reducing FLOPs by about 44% compared to MedViT variants.

Beyond generic classification benchmarks, attention-augmented Transformers have been tailored to specific clinical modalities. Med-Former [116] combines a Local–Global Transformer (LGT) module with a Spatial Attention Fusion (SAF) module to jointly capture multi-scale contextual information and propagate salient features across stages, thereby enhancing the feature extraction capability of Transformer backbones for medical images. On National Institutes of Health (NIH) Chest X-ray14 (a large-scale chest X-ray dataset with 14 thoracic disease labels), Med-Former attains 87.6% AUC for multi-label thoracic disease classification and also achieves competitive results on DermaMNIST (a MedMNIST-2D dermatoscopic skin lesion classification dataset) and BloodMNIST (a MedMNIST-2D blood cell microscopy dataset for 8-class blood cell type classification), with up to 96.5% accuracy and 99.7% AUC, respectively. For fundus disease screening, SwinECAT [89] couples shifted-window self-attention with an Efficient Channel Attention (ECA) module, guiding the model to focus on lesion-relevant channels while preserving the hierarchical Swin Transformer structure. On the Eye Disease Image Dataset (EDID), SwinECAT achieves 88.29% accuracy for nine-class fundus disease classification, substantially outperforming the Swin Transformer baseline.

In the context of multi-modality disease analysis, CDDFuse [91] introduces a correlation-driven dual-branch Transformer–CNN feature decomposition network that disentangles low-frequency modality-shared features and high-frequency modality-specific details. By combining Lite Transformer blocks for global context, invertible neural networks for detail preservation, and a correlation-driven loss, CDDFuse produces fused infrared–visible and medical images that significantly boost downstream semantic segmentation and lesion detection performance compared with the previous fusion methods.

For multi-label medical image classification, CTransCNN [90] proposes a parallel CNN–Transformer architecture in which the Multi-Modal Attention Enhancement Fusion (MMAEF) module mines label correlations, the Model Balance Refinement (MBR) module mitigates data imbalance, and the Inter-Interaction Module (IIM) facilitates bidirectional information exchange between CNN and Transformer branches. On ChestX-ray11 (a large-scale chest X-ray dataset with 11 thoracic disease labels) and NIH ChestX-ray14, CTransCNN attains average AUCs of 83.37% and 78.47%, respectively, and reaches 84.56% AUC on the TCMTD [90] tongue dataset, outperforming strong CNN and pure Transformer baselines for rare-label prediction. These improvements come at the expense of more complex training objectives and heavier backbones, which may increase the deployment cost in routine clinical screening.

Table 3 summarizes representative attention- and Transformer-based models for medical image classification and segmentation. In addition to accuracy (ACC), DSC, and AUC on public benchmarks, the table reports the key hardware configurations (as stated in the original papers) and highlights the main architectural advantages and limitations. Overall, these methods exhibit clear accuracy–complexity trade-offs: HiFuse-Base attains strong skin-lesion classification performance on ISIC2018 with a relatively heavy backbone of about 82.49 M parameters and 8.13 GFLOPs; MedViT-S and MedViTV2-S maintain high AUC on the MedMNIST family while explicitly optimizing parameter/FLOP efficiency, with MedViTV2-S reducing computational complexity by approximately 44% compared with MedViT. Med-Former and CTransCNN further leverage multi-head self-attention and label-correlation modeling to improve multi-label chest X-ray (i.e., classifying multiple thoracic findings from frontal chest radiographs) AUC, but introduce more complex training objectives and heavier backbones than conventional CNNs. In the segmentation domain, SCSE-enhanced DenseNet [85] achieves competitive Dice scores on brain MRI yet still relies on dense voxel-level supervision and 3D convolutional encoders.

From a clinical perspective, these models target complementary application scenarios. HiFuse and MedViT/MedViTV2 are well-suited for single-image disease classification and large-scale screening on standardized datasets, where offline training on high-end GPUs (RTX 3090, 2080Ti, A100) is acceptable and inference can be batched. Med-Former and CTransCNN are designed for multi-label thoracic disease

recognition and catheter assessment on chest radiographs, providing richer multi-disease predictions, but at a higher computational cost, which is more appropriate for server-side computer-aided diagnosis systems than for edge deployment. SwinECAT [89] focuses on the fundus disease screening and shows that lightweight Swin-based attention can effectively highlight fine-grained retinal lesions, although its generalization beyond EDID remains to be verified. SCSE-enhanced DenseNet mainly addresses 3D brain structure segmentation and demonstrates that plugin-style attention blocks can improve volumetric delineation without redesigning the entire backbone. Taken together, attention and Transformer modules substantially enhance recognition and segmentation performance in diverse medical imaging tasks, but their real-world clinical adoption still depends on achieving a suitable balance between predictive accuracy, model complexity, and hardware constraints.

Table 3: A summary of representative models for medical image classification and recognition.

Model	Dataset	Task	ACC	DSC	AUC	Hardware	Advantages	Limitations	Application	Ref.
HiFuse-Base	ISIC 2018	Multi-class	85.85	–	95.85	RTX 3090	Three-branch hierarchical fusion aggregates local and global features effectively.	Heavy backbone (82.49M params); costly training compared to CNNs.	Dermoscopic skin-lesion classification.	[86]
MedViT-S	Med-MNIST	Multi-task	85.10	–	94.20	RTX 2080Ti	Hybrid framework modeling local detail and global context; strong generalization.	Training from scratch on small datasets is costly; relies on regularization.	General-purpose medical image classification.	[87]
Med-Former	Chest X-ray14	Multi-label	–	–	87.60	V100	Jointly captures multi-scale context; competitive multi-label performance.	Long-range dependencies increase inference time; limited prospective evaluation.	Multi-disease chest X-ray screening.	[116]
MedViT-V2-S	Path-MNIST	Multi-class	96.50	–	99.80	A100	KAN-integrated Transformer with Dilated Neighborhood Attention (DiNA); reduces FLOPs by 44% over MedViT.	Requires high-end GPUs; complex architecture hinders simple re-implementation.	Robust classification under corrupted conditions.	[88]
Swin-ECAT	EDID	Multi-class	88.29	–	–	Not reported	Combines Swin Transformer with Efficient Channel Attention for small lesions.	Limited cross-modality validation; hardware details not fully reported.	Fundus disease classification.	[89]

(Continued)

Table 3 (continued)

Model	Dataset	Task	ACC	DSC	AUC	Hardware	Advantages	Limitations	Application	Ref.
CTrans-CNN	Chest X-ray11	Multi-label	–	–	83.37	A5000	Explicitly models label correlations and cross-branch interactions.	Complex training objectives; heavier backbone than pure CNNs.	Thoracic disease classification.	[90]
SCSE-DenseNet	MALC	Seg/Class	–	88.20	–	Titan Xp	Recalibrates spatial/channel responses; improves segmentation accuracy.	Relies on voxel-level supervision; limited to 3D CNN backbones.	Brain MRI structure segmentation.	[85]

Note: Units: ACC, DSC, AUC in %. Values are based on original papers. "–" indicates metric not reported.The "Ref." column cites the original paper introducing each model.

3.2.2 Applications in Few-Shot and Weakly Supervised Scenarios

Medical image annotation relies heavily on expert knowledge and is both costly and time-consuming, which leads to limited labeled datasets in many clinical domains. Consequently, few-shot learning (FSL) and weakly supervised learning have become important research directions. Attention-augmented networks already show clear advantages in these data-scarce regimes, and the introduction of Transformer architectures further amplify their potential. In FSL settings, the ability to model long-range dependencies and to reuse features across related tasks are crucial. Transformer-based segmentation models such as TransUNet [65] and Swin-UNet [17], although originally designed for fully supervised learning, have been shown to outperform purely convolutional baselines when trained with reduced annotation budgets. Their hybrid CNN–Transformer encoders exploit both local texture cues and global anatomical context, which helps maintain stable Dice and AUC scores even when only a small fraction of the training images are labeled. However, these models typically require substantially higher GPU memory and longer training time than lightweight CNNs, which may limit their adoption in smaller hospitals and resource-constrained clinical environments. In weakly supervised scenarios, models are trained with sparse or incomplete labels such as image-level tags, coarse scribbles, or partial annotations. Self-attention provides an implicit mechanism for focusing on salient structures under such supervision. MedT [40] introduces gated axial attention and a local–global (LoGo) training strategy that improves data efficiency and allows Transformers to be trained from scratch on relatively small medical datasets. UNETR [69] replaces the convolutional encoder in a 3D U-Net with a ViT-based encoder, enabling the model to capture volumetric long-range dependencies and to delineate complex anatomical structures, even when the number of labeled volumes is limited. Beyond fully supervised architectures, explicitly weakly supervised methods further exploit attention mechanisms. SA-MIL [117] incorporates self-attention into a multiple instance learning framework for histopathology, modeling global correlations among image patches and using deep weak supervision to obtain pixel-level tumor segmentation from slide-level labels. SSL-Transformer [118] introduces a cross-teaching framework between a CNN and a Transformer, utilizing the Transformer's self-attention to capture global contexts and enforce prediction consistency with the CNN, thereby achieving robust semi-supervised segmentation performance on cardiac and abdominal datasets. Taken together, these studies indicate that attention-augmented CNNs and Transformer-based architectures can substantially improve robustness and generalization under both few-shot and weakly supervised conditions. At the same time, their higher computational cost, increased memory footprint, and sensitivity to training hyperparameters highlight

the need for more efficient and clinically deployable designs, as well as for rigorous benchmarking on standardized low-annotation medical datasets.

3.3 Object Detection and Localization

3.3.1 Attention-Enhanced Detection Frameworks

Lesion detection and localization in medical images is particularly challenging when the targets are small, low-contrast, or embedded in complex anatomical backgrounds. Conventional two-stage detectors such as Faster R-CNN (Region-based Convolutional Neural Network) suffer from proposal sparsity and scale mismatch, which limit sensitivity to micro-lesions and blurred boundaries. To address these issues, recent works integrate attention mechanisms and Transformer architectures into both proposal-based and end-to-end detectors, and redesign anchors and feature pyramids to better capture tiny lesions. For proposal-based methods, attention-enhanced Mask R-CNN variants [119] incorporate spatial-channel attention and deformable convolutions. This design effectively strengthens feature representation for geometric variations, significantly outperforming the vanilla baseline in detecting small brain CT lesions. End-to-end detectors such as EL-DETR (Detection Transformer) [92] adopt CNN–Transformer hybrids with multi-head self-attention and query-based decoding to model the global context and obtain high-precision lesion detection without hand-crafted anchors. PMA-DETR [93] further incorporates multi-view and dilated convolutions together with a pyramid multi-level attention module, significantly improving mAP on Breast Ultrasound Images (BUSI) breast ultrasound and brain tumor benchmarks compared with Faster R-CNN, RetinaNet (introduced in [120]), and vanilla DETR. Recent one-stage frameworks emphasize efficiency and anchor design for small lesions. GravityNet [94] replaces fixed anchor boxes with pixel-based "gravity points" that are iteratively attracted to lesion candidates, yielding competitive Free-response Receiver Operating Characteristic (FROC) performance for microcalcification and microaneurysm detection under strong class imbalance. MedYOLO [95] extends the YOLO (You Only Look Once) family to 3D medical volumes by redesigning the detection head and anchor generation for volumetric inputs; on BRaTS [121], LIDC [122], and abdominal CT datasets it achieves high mAP for medium and large structures while retaining single-shot efficiency, though performance still decreases for extremely small or rare lesions. Overall, attention-augmented two-stage and Transformer-based detectors jointly improve small-lesion sensitivity and localization accuracy, but also increase computational cost and GPU memory requirements.

3.3.2 Segmentation-Oriented Micro-Lesion Localization

For subpixel-level localization and boundary refinement, many attention-based networks formulate lesion detection as segmentation, and then derive bounding boxes from connected components. CaraNet [80] introduces contextual-axis reverse attention and a channel feature pyramid to progressively suppress easy background and emphasize small objects, achieving a top-ranked mean Dice on five polyp datasets and BraTS 2018 tumor segmentation, especially when the lesion areas are smaller than 5% of the image. SvANet [81] employs a scale-variant attention strategy with cross-scale guidance and stochastic multi-scale fusion to better capture tiny structures such as kidney tumors, skin lesions and vessels across seven heterogeneous datasets. STS-Net [82] focuses on extremely small, sparse lesions by combining lesion amplification blocks with an encoder-side enhanced spatial–channel attention module. Typical target lesions include cavernous malformations and ultrasound micro-lesions. On SCCM-2022 [82] and Breast Ultrasound Images Dataset – Small-lesion subset (BUID-S), it consistently outperforms U-Net, nnU-Net, and DeepLabv3+ in terms of Dice and Mean Intersection over Union (mIoU) while remaining deployable on mid-range GPUs [82].More recent attention-centric architectures such as MCANet [83] and SMAFormer [84] integrate multi-scale attention designs with Transformer-style feature modeling, achieving

strong performance on multi-organ and lesion benchmarks and improving the delineation of irregular and fuzzy boundaries. These methods highlight that carefully designed attention modules are essential for robust micro-lesion localization under severe class imbalance and low contrast.

From an implementation perspective, most of the detectors and segmentation networks in Tables 4 and 5 are trained on single modern GPUs, but only a subset of works explicitly report the type of accelerator or memory capacity. Two-stage frameworks and DETR-style detectors rely on multi-scale backbones and global self-attention, which increases parameter counts, FLOPs, and GPU memory compared with classical Faster/Mask R-CNN; representative DETR-style variants include EL-DETR, PMA-DETR, and one-stage CNNs. 3D extensions such as MedYOLO further exacerbate memory pressure when applied to full-resolution CT/MR volumes. In contrast, GravityNet and lightweight YOLO-style designs offer faster inference and more favorable computational profiles, but they may show reduced sensitivity to extremely small or diffuse lesions. For micro-lesion-oriented segmentation, CaraNet provides a relatively efficient convolutional baseline, whereas SvANet and STS-Net use richer attention blocks and lesion amplification strategies to improve Dice and mIoU at the cost of more complex training and higher Video Random Access Memory (VRAM) requirements. From a clinical perspective, proposal-based detectors and DETR variants are well suited to high-precision polyp and lesion mining in colonoscopy, breast ultrasound and CT screening, where interpretable bounding boxes and attention maps can guide radiologists toward suspicious regions. Segmentation oriented architectures such as CaraNet, SvANet, and STS-Net are preferable when sub-pixel boundary delineation or volumetric burden estimation is critical, for example, in small liver/kidney tumors, cavernous malformations or early micro-lesions in ultrasound. Nevertheless, most studies are still validated on public or single-center datasets, and prospective multi-center trials are rare. Overall, attention and Transformer modules significantly improve small-lesion sensitivity and localization, but their clinical translation hinges on achieving a robust balance between accuracy, computational complexity, and ease of deployment in routine Computer-Aided Diagnosis (CAD) pipelines.

Table 4: A summary of representative attention-based models for medical lesion detection.

Model	Dataset	Task	mAP_{50}	Other	Hardware	Advantages	Limitations	Application	Ref.
Mask R-CNN+	Brain CT	Lesion Det	–	98.1 (ACC)	RTX 2080Ti	Deformable convs + attention enhance small-lesion sensitivity.	Two-stage cascade increases inference time; complex parameter tuning.	Brain lesion grading and detection.	[119]
EL-DETR	Polyps	End-to-end	98.0	71.0 (mAP_{95})	RTX 3090	Global self-attention with query-based decoding; interpretable maps.	Requires long training schedules and substantial GPU memory.	High-precision polyp detection.	[92]
Gravity-Net	Mammo	Micro-calc	–	72.3 (FROC)	RTX-series	Pixel-based "gravity point" anchors adapt to tiny targets.	Specialized for 2D tiny lesions; generalization to 3D is less explored.	Screening for microcalci-fications.	[94]
Med-YOLO (3D)	BRaTS	3D Det	86.1	43.1 (mAP_{95})	–	Single-shot 3D detection; efficient for medium/large structures.	3D convolutions increase memory usage; accuracy drops for tiny lesions.	Volumetric tumor/organ detection.	[95]

Note: Units in % unless otherwise stated. Mask R-CNN+ refers to the attention-enhanced variant. Ref. denotes the original model paper (same convention as Table 3).

Table 5: A summary of representative attention-based segmentation models for micro-lesion localization.

Model	Dataset	Task	DSC	mIoU	Hardware	Advantages	Limitations	Application	Ref.
CaraNet	Kvasir	Polyp Seg	94.0	90.0	RTX-series	Contextual-axis reverse attention strengthens small-object features.	Purely convolutional; limited global context modeling.	Small polyp and brain tumor segmentation.	[80]
SvANet	KiTS	Small Seg	89.8	–	RTX 4090	Scale-variant attention robust to large shape/scale variations.	Complex attention blocks increase FLOPs and memory usage.	Kidney, skin, and vessel segmentation.	[81]
STS-Net	SCCM	Tiny Seg	82.1	77.5	RTX 3090	Lesion amplification and enhanced attention boost tiny lesion signal.	Augmentation strategy may introduce artifacts if applied blindly.	Cavernous malformations and micro-lesions.	[82]

Note: Units: DSC and mIoU in %. "–" indicates metric not reported. Ref. denotes the original model paper (same convention as Table 3).

3.4 Image Registration

Medical image registration aims to align images acquired at different times, from different modalities, or across individuals into a common spatial coordinate system. This process underpins structural comparison, longitudinal lesion tracking, and multimodal fusion. Classical optimization-based registration methods depend on handcrafted features and similarity metrics, which are computationally expensive and sensitive to initialization. CNN-based frameworks such as VoxelMorph greatly improve inference efficiency, but their local receptive fields limit the modeling of long-range anatomical correspondences and complex 3D deformation fields. To address these limitations, recent works integrate attention mechanisms and Transformer architectures into registration networks, enabling explicit global dependency modeling in both image and deformation spaces. In what follows, representative attention- and Transformer-based registration and related cross-modal alignment frameworks are briefly reviewed.

3.4.1 Attention-Augmented CNN Registration Networks

Within the VoxelMorph paradigm, DaVoxelMorph [96] introduces a dual attention architecture that explicitly models correlations in both spatial and coordinate dimensions, so that each voxel aggregates information from all spatial locations while preserving deformation regularity. A location attention module weights global spatial features, whereas a coordinate attention module injects positional cues into channel attention to enhance structural consistency. On the LPBA40 brain MRI dataset, DaVoxelMorph [96] yields superior Dice scores and topology preservation compared to the baseline VoxelMorph, while maintaining efficient training on mid-range GPUs. Similar attention-augmented CNN variants further employ multi-scale correlation constraints and spatial/channel attention to better preserve small anatomical structures under large deformations, but their receptive fields remain inherently limited by convolution.

3.4.2 Transformer-Based Volumetric Registration

TransMorph [97] is a hybrid Transformer–ConvNet framework that replaces the encoder of VoxelMorph with a 3D Swin Transformer, substantially enlarging the effective receptive field and enabling more accurate voxel-wise correspondence modeling between moving and fixed images. The authors also propose diffeomorphic and Bayesian variants that produce topology-preserving deformations and calibrated uncertainty estimates. On Open Access Series of Imaging Studies (OASIS), Information eXtraction from Images (IXI) and Learn2Reg brain MRI benchmarks, as well as phantom-to-patient CT registration, TransMorph and its variants consistently outperform CNN-based baselines in the Dice score and deformation smoothness while remaining trainable on TITAN RTX/RTX 3090 GPUs. Building on such hybrid designs, several recent works (2023–2025) directly enhance global modeling in the deformation field. XMorpher [123] presents a full Transformer architecture for unsupervised deformable registration. It utilizes a dual-stream encoder to extract features from moving and fixed images separately and employs a cross-attention module to explicitly model the non-linear correspondence between diverse anatomical features, achieving state-of-the-art registration accuracy on brain MRI datasets. Deformable Cross-Attention (DCA) Transformers relax the fixed-window constraint of standard cross-attention by allowing queries to sample from a learnable, potentially image-wide search region, thus capturing long-range correspondences with manageable computational cost [124]. In parallel, diffusion-based Transformers such as DiffuseReg [98] formulate unsupervised registration as denoising of a deformation field within a Swin-Transformer-based diffusion model, improving Dice scores and Jacobian regularity on cardiac MRI at the price of longer inference time [98]. Overall, these Transformer-based frameworks show that explicitly modeling long-range interactions in image or deformation space significantly boost accuracy and stability, especially in challenging 3D neuroimaging scenarios.

3.4.3 Cross-Modal and Semantically Guided Registration

In multi-modal scenarios, attention is used to align heterogeneous feature spaces and inject semantic priors into the registration process. Representative cross-modal registration tasks include CT–MR fusion and image–text alignment. MedKLIP [125] employs a triplet-wise attention mechanism to incorporate medical domain knowledge. It decouples the image-text matching process by using attention to align visual features with distinct semantic queries derived from medical knowledge bases, significantly improving zero-shot diagnosis performance. CycleMorph [126] incorporates a cycle-consistent attention mechanism to align MRI and CT images. By enforcing geometric consistency through a cycle constraint and focusing on anatomically shared features via attention, it effectively handles the intensity heterogeneity between modalities without requiring ground-truth deformation fields. MGCA [109] further proposes multi-granularity cross-modal alignment that matches medical images and radiology reports at the token, instance, and prototype (disease) levels via cross-attention, yielding stronger medical visual representations. Although these models mainly target vision–language understanding instead of voxel-level geometric registration, they demonstrate how cross-attention can build shared latent spaces across modalities. CMiNet (Cross-Modality Interaction Network) extends this idea to medical image fusion via a recursive Transformer that captures long-range intra-modality context and an interactive CNN fusion module, producing fused images with improved anatomical and functional detail [127]. Such cross-modal attention designs provide useful priors for future multi-modal registration, for example, by using fused images or text descriptions to constrain the deformation search space in data-limited clinical settings.

Table 6 summarizes representative attention- and Transformer-based registration frameworks and highlights their hardware requirements and limitations. In general, attention-augmented CNNs such as DaVoxelMorph provide moderate Dice and Jacobian improvements over VoxelMorph with relatively small overhead and can be trained on mid-range GPUs, making them suitable for routine 3D brain MRI registration on standard workstations. Transformer-based models such as TransMorph and H-ViT substantially expand the receptive field and achieve higher Dice scores and smoother deformation fields across OASIS, IXI, and related benchmarks, but their self-attention blocks lead to increased parameter counts, FLOPs, and memory usage, and typically require high-end GPUs and longer training time. In practice, such models are typically trained on high-end accelerators such as TITAN RTX and RTX 3090. Diffusion-based approaches like DiffuseReg further improve topology preservation and robustness by iteratively denoising the deformation field, yet incur the highest inference latency among the reviewed methods. Overall, attention mechanisms and Transformer architectures have become central to modern medical image registration by enabling multi-scale global dependency modeling and semantic cross-modal alignment. At the same time, most evaluations are still conducted on public neuroimaging and cardiac datasets under retrospective settings, and direct clinical deployment will require more efficient model designs, rigorous validation on prospective multi-center cohorts, and careful balancing of registration accuracy, computational complexity, and hardware constraints in real-world clinical workflows.

Table 6: A summary of models for deformable medical image registration.

Model	Dataset	Task	DSC	Hardware	Advantages	Limitations	Application	Ref.
DaVoxel- Morph	LPBA40	Unsup 3D	71.4	RTX A2000	Dual attention improves topology preservation with low cost.	Still constrained by convolutional receptive fields; moderate gains.	Fast 3D brain MRI registration.	[96]
Trans- Morph	OASIS	Volume-tric	–	Titan RTX	Hybrid Swin-Transformer encoder greatly enlarges receptive field.	Higher parameter count and memory footprint than CNNs.	High-accuracy inter-patient registration.	[97]
Diffuse- Reg	Cardiac	Diffu-sion	–	–	Diffusion-based denoising improves deformation regularity.	Iterative process introduces substantial inference overhead.	Offline car-diac/abdominal registration.	[98]
DCA- Trans	Brain	Deform-able	–	–	Deformable Cross-Attention captures long-range cor-respondences.	Complex sampling increases difficulty; heavier than CNNs.	Large non-linear deformation registration.	[124]

Note: DSC in %. Unsup = Unsupervised. Ref. denotes the original model paper (same convention as Table 3).

3.5 Image Generation and Enhancement

Medical image generation and enhancement aim to learn high-dimensional feature distributions and synthesize or restore images with realistic anatomical structures and textures, thereby supporting diagnosis, data augmentation, and model generalization. Because annotated medical data are expensive to acquire, especially for rare diseases and specific modalities, generative models based on Generative Adversarial Networks (GANs) and diffusion processes have become important tools to alleviate data scarcity and domain shift. The integration of attention mechanisms and Transformer architectures further strengthens global

context modeling and fine-grained structural reconstruction. In this subsection, representative attention- and Transformer-based frameworks for medical image generation and enhancement are briefly reviewed, highlighting their core architectural designs and typical application scenarios.

3.5.1 Attention-Augmented GANs

DAGAN [99] incorporates a position-wise spatial attention module into the generator and discriminator to suppress artifacts in MRI reconstruction, ensuring that the model focuses on restoring anatomical details rather than background noise. Furthermore, pure Transformer-based architectures have been adapted for medical synthesis; for example, SwinGAN [100,101] leverages the hierarchical Swin Transformer to synthesize high-fidelity medical images (e.g., PET-to-CT translation). Its shifted-window attention mechanism effectively models long-range dependencies in volumetric data, achieving superior structural consistency compared to CNN-based GANs.

3.5.2 Transformer-Based Restoration and Diffusion Models

For image restoration, SwinMR [102] adapts the hierarchical Swin Transformer for fast MRI reconstruction. By jointly modeling local textures and global structures via shifted-window attention, it preserves organ boundaries and subtle lesion details more effectively than conventional convolutional approaches.

Diffusion models improve generative quality by iteratively denoising noisy inputs. Recently, MedSegDiff [128] adapts the diffusion transformer architecture for medical image segmentation and generation, utilizing an attention-based conditioning mechanism to integrate multi-modal constraints. This approach demonstrates superior performance in capturing anatomical ambiguity compared to traditional U-Net discriminative models. DPM-MedImgEnhance [129] trains diffusion models on high-quality CT and cardiac MR data, and uses them as plug-and-play priors for low-dose CT denoising and MR super-resolution without requiring paired training samples, achieving higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) than conventional reconstruction pipelines. UniMIE [103] leverages a pre-trained natural-image diffusion model in a training-free manner and achieves universal enhancement across 13 modalities and 15 tasks via a learned degradation mask and exposure-control losses, which greatly reduces the need for modality-specific retraining. For volumetric data, 3D MedDiffusion [104] combines a 3D latent autoencoder with a diffusion process to synthesize and enhance high-resolution 3D CT and MR volumes, supporting sparse-view reconstruction and data augmentation for downstream segmentation and detection networks.

Compared with GAN-based approaches, diffusion models generally offer more stable training and stronger generative fidelity, but the iterative sampling procedure significantly increases inference time and often requires high-memory GPUs, especially in 3D settings. In addition, both GAN and diffusion models may hallucinate anatomically plausible but clinically incorrect structures or alter the subtle lesion appearance, which poses safety risks if enhanced images are used directly for diagnosis. Most studies remain confined to retrospective single- or multi-center datasets, and prospective clinical validation is still limited. Overall, attention and Transformer modules enhance global context modeling and structural fidelity for medical image generation and enhancement, yet practical adoption in clinical workflows will depend on carefully balancing image quality improvements against computational complexity, memory requirements, interpretability, and the need for rigorous clinical validation.

3.6 Multimodal Data Fusion

Multimodal medical image fusion aims to integrate complementary information from different imaging modalities to obtain more comprehensive representations of anatomical structures and functional characteristics. Typical modalities considered in fusion studies include CT, MRI, and PET. Structural modalities emphasize high-resolution morphology, whereas functional modalities capture metabolic or perfusion patterns; effective fusion, therefore, requires both accurate cross-modal alignment and task-aware feature integration. Classical pixel-level fusion or simple feature concatenation is often insufficient to model long-range semantic correspondences across modalities. Recent attention and Transformer architectures address these limitations by explicitly learning cross-attention maps between modalities and by decoupling shared vs. modality-specific representations.

3.6.1 Cross-Attention–Based Multimodal Image Fusion

TransMed [105] treats multi-modal MRI slices as token sequences, utilizing self-attention to capture long-range inter-modal dependencies. This design significantly improves diagnostic accuracy for parotid tumors and knee injuries compared to pure CNN baselines. For low-level structural–functional fusion, CDDFuse [91] decomposes cross-modal features into shared low-frequency and modality-specific high-frequency components using a dual-branch Transformer–CNN backbone: Lite Transformer blocks model global low-frequency context while invertible neural network blocks preserve high-frequency texture, guided by a correlation-driven constraint on the two branches. MATR [106] proposes a multi-scale adaptive Transformer for multimodal medical image fusion. It utilizes a cross-attention mechanism to align and fuse features from different modalities, adaptively emphasizing functional information from PET while retaining the structural details from MRI, thereby enhancing tumor visibility. Beyond purely spatial-domain designs, recent diffusion-based fusion models such as DM-FNet [107] embeds cross-modal fusion modules into the denoising process, learning a generative prior over multimodal brain images and achieving high-quality MRI–CT, MRI–PET, and MRI–single-photon emission computed tomography (SPECT) fusion while maintaining detailed anatomy and functional contrast. These methods illustrate a trend from handcrafted fusion rules toward learned cross-attention or diffusion priors that jointly optimize structural fidelity, contrast, and downstream task performance.

3.6.2 Vision–Language Fusion for Image–Text Understanding

Vision–language pre-training extends multimodal fusion from images alone to joint image–text representation learning. MedCLIP [108] adapts CLIP(Contrastive Language-Image Pre-training)-style contrastive learning to medical data by aligning unpaired images and reports through a semantic entity matrix and text augmentation, enabling zero-shot classification, retrieval, and report matching under limited supervision. MGCA [109] further exploits multi-granularity cross-modal alignment at the region, instance, and disease levels with multiple cross-attention branches, achieving strong transfer to diverse chest X-ray and CT benchmarks by explicitly modeling correspondence between visual regions and textual findings. BioViL-T [110] incorporates temporal context from current and prior chest radiographs and their associated reports, using self-attention within each modality and cross-attention across modalities and time to capture disease progression patterns; this yields state-of-the-art performance for classification, report generation, and phrase grounding in radiology. Recent work such as HCFNet [130] shows that hybrid cross-modality fusion decoders, coupled with multi-stage image–text contrastive objectives, can enhance multimodal segmentation and provide more reliable lesion delineation in scenarios where aligned images and clinical text are both available. Overall, cross-attention–driven multimodal fusion and vision–language modeling provide a key

route from single-modality perception toward the semantic-level understanding and cross-modal reasoning in medical AI.

From a computational perspective, these multimodal fusion and vision–language frameworks also exhibit diverse complexity profiles. Slice-level fusion networks such as TransMed and CDDFuse typically rely on CNN or hybrid CNN–Transformer backbones with dual-branch encoders and cross-attention modules, which introduce additional parameters and FLOPs compared with single-modality CNNs but remain trainable on single high-memory GPUs. In contrast, diffusion-based fusion models like DM-FNet and large-scale vision–language pre-training frameworks such as MedCLIP, MGCA, and BioViL-T require long training schedules on large multimodal corpora and often depend on multi-GPU or high-VRAM accelerators, which raises the barrier for routine deployment in smaller institutions.

From a clinical perspective, cross-attention-based image fusion models have demonstrated improvements in contrast, edge preservation, and downstream segmentation or detection performance on public PET/CT, MRI/CT, and infrared–visible benchmarks, while vision–language models achieve strong zero-shot or few-shot performance on radiographic classification, report generation and phrase grounding tasks. However, most evaluations are still retrospective and limited to a small number of curated datasets; systematic prospective validation across institutions and imaging protocols is scarce. Moreover, complex fusion and vision–language architectures can inadvertently amplify modality-specific artefacts or hallucinate spurious cross-modal correspondences, which raises safety concerns if their outputs are used directly for diagnosis or treatment planning. Overall, attention- and Transformer-based multimodal fusion methods offer clear advantages in modeling long-range semantic correspondences and leveraging complementary information across modalities, but their clinical adoption will depend on carefully balancing performance gains against computational cost, interpretability, and the need for rigorous, task-specific clinical verification.

3.7 Summary and Critical Analysis of Applications

The application of attention mechanisms in medical image analysis has evolved from simple feature enhancement to serving as the foundational architecture for global context modeling. Across the diverse tasks discussed in Sections 3.1–3.6, three distinct evolutionary trajectories and trade-offs that define the current state of the field can be identified.

Architectural Evolution: From Enhancement to Hybridization. Early approaches utilized attention as a lightweight "plugin" to suppress background noise in CNNs. While effective for local feature refinement, these methods failed to capture the long-range dependencies required for understanding complex anatomical structures. This limitation drove the shift towards pure Transformers. However, recent trends favor hybrid architectures. Analysis suggests that hybrid designs offer an optimal compromise for medical imaging: CNN layers preserve the high-frequency spatial details crucial for boundary delineation that pure Transformers often lose, while attention layers provide the global semantic shape consistency that CNNs lack.

Task-Specific Mechanisms and Pain Points. Different medical tasks leverage attention to solve fundamentally different problems. In segmentation and registration, the primary challenge is topological consistency. Self-attention addresses this by modeling global voxel-to-voxel correlations, enabling models to handle large deformations and discontinuous organ shapes that limited-receptive-field convolutions cannot resolve. In detection, the focus is on "rare event" discovery. Cross-attention mechanisms function as dynamic scanning tools, breaking the limitations of fixed anchors to detect extremely small or sparse lesions in large volumetric scans. Regarding multimodal fusion, the core issue is the "semantic gap" between modalities. Cross-attention serves as a learnable alignment bridge, weighting information from one modality conditioned on the reliability of another, significantly outperforming naive concatenation.

The Efficiency-Accuracy Trade-off. While attention mechanisms consistently improve metrics such as DSC, AUC, and mAP, they introduce substantial computational overhead. Global self-attention's quadratic complexity ($O(N^2)$) remains a bottleneck for 3D medical volumes. Consequently, the field is moving away from global dense attention towards efficient variants—such as window-based attention, sparse attention, and linear complexity approximations—to make these advanced models deployable in resource-constrained clinical environments.

4 Challenges and Future Research Directions

Although attention mechanisms have been widely adopted in medical image analysis, many challenges remain and warrant further research. Based on the emerging needs of real-world clinical applications, several promising research directions are outlined in this section. This review provides a comprehensive and cutting-edge analysis of attention mechanisms in medical image analysis, offering a unified multi-perspective classification framework and critical performance benchmarking that distinguishes it from existing surveys. The systematic comparison of attention-based models against traditional architectures, coupled with an in-depth analysis of emerging trends such as large-scale foundation models and federated learning, provides researchers with actionable insights and forward-looking guidance.

4.1 Current Challenges

4.1.1 Comparative Analysis: Attention vs. Non-Attention Deep Models

While attention mechanisms offer significant advantages in capturing long-range dependencies and adaptive feature re-weighting, it is essential to critically compare them with traditional non-attention deep models, including CNNs, Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs). CNNs excel at capturing local spatial patterns through convolutional filters but struggle with global context integration. RNNs handle sequential data effectively but face challenges with long-term dependencies. GNNs model relational structures but require predefined graph topology. In contrast, attention mechanisms dynamically focus on relevant regions across the entire input, enabling better interpretability and handling of complex dependencies. However, this comes at the cost of higher computational complexity and data requirements compared to lightweight CNNs. The persistence of these trade-offs underscores the need for hybrid architectures that combine the efficiency of CNNs with the global reasoning capability of attention.

4.1.2 Medical Data Scarcity

In real-world clinical settings, the scarcity of medical data is a common issue. Data acquisition and annotation are costly and time-consuming, which discourages healthcare institutions from undertaking these tasks. In addition, for rare diseases, it is inherently difficult to collect large-scale datasets. However, when training attention-based image recognition models, limited data can easily lead to overfitting. This makes it difficult for the model to capture key features of the lesion. Furthermore, the integration of attention mechanisms with transfer learning and self-supervised learning remains underdeveloped, which limits the model's generalization capability. This challenge persists due to privacy regulations, annotation complexity, and the domain-specific nature of medical imaging that limits transferability from natural image datasets. Potential solutions include developing specialized data augmentation techniques for medical images and creating standardized benchmarking datasets.

4.1.3 Insufficient Generalization and Cross-Center Adaptability

In medical imaging, differences in devices and imaging protocols can cause domain shifts. These shifts often lead to poor performance of trained models in practical tasks. However, traditional attention mechanisms do not respond consistently to data from different centers. Their feature re-weighting methods also lack domain adaptability. As a result, the generalization and transfer capacity of the model remain limited, creating challenges for clinical deployment. The fundamental issue lies in the assumption of independent and identically distributed data during training, which rarely holds in multi-center medical studies. Recent advances in domain generalization and test-time adaptation offer promising directions, though their integration with attention mechanisms remains limited.

4.1.4 Complexity of Multi-Modal Feature Fusion

The use of various medical imaging techniques—such as CT, MRI, PET, and ultrasound—provides richer information for diagnosis but also introduces challenges in fusing multi-modal data. Different types of medical images have different feature distributions, making semantic alignment between modalities difficult. Although cross-attention mechanisms can enable multi-modal fusion to some extent, they face computational and fusion bottlenecks when processing high-dimensional medical data. In addition, complementary information between modalities is often underutilized, which limits the effectiveness of fusion. Case studies from recent literature demonstrate that effective multi-modal fusion can improve diagnostic accuracy by 15%–20% compared to single-modality approaches, though quantitative benchmarks remain limited [131].

4.1.5 Model Interpretability and Clinical Trustworthiness

Although attention mechanisms can focus on key regions annotated in medical images, this process often lacks a clear physiological or anatomical basis and relies heavily on the accuracy of manual annotations. As a result, the medical interpretability of attention remains limited, making it difficult for clinicians to understand the reasoning behind its decisions. Specifically, while attention maps highlight relevant regions, they often fail to provide causal explanations for why certain regions are emphasized. Integration with established visualization techniques like Grad-CAM and saliency maps could enhance interpretability [132], though these methods themselves have limitations in medical contexts. In addition, attention weights can vary significantly with slight changes in input data, leading to unstable and non-reproducible model behavior—which is unacceptable in medical practice.

4.1.6 High Computational Costs and Barriers to Clinical Deployment

The deployment of attention-based models on clinical equipment faces the challenge of high computational complexity. On the one hand, Transformer-based architectures inherently contain a large number of parameters, leading to high computational costs during both training and inference. On the other hand, medical images—particularly high-resolution 3D sequences such as CT scans—involve extremely large datasets, which further significantly increases the computational load of attention mechanisms. However, in reality, clinical devices generally have limited computational power, making it difficult to meet the requirements of current attention-based models. As a result, the deployment of such models remains restricted. Quantitative analysis shows that standard Transformer attention scales quadratically with input size, making high-resolution 3D medical images computationally prohibitive [13].

4.1.7 Data Shift and Robustness Risks

Although attention mechanisms can focus on key regions in images, this may also cause the model to overemphasize certain features, leading to model bias. In addition, attention mechanisms often lack robustness against common medical imaging issues such as noise, artifacts, and abnormal samples. Furthermore, they usually lack mechanisms for uncertainty modeling and confidence estimation in predictions. This challenge is particularly acute in medical imaging where acquisition parameters, patient populations, and imaging protocols vary widely across institutions.

4.2 Future Directions for Attention Mechanisms

4.2.1 Self-Supervised and Few-Shot Attention Learning

The scarcity of medical data is largely due to the high cost and effort required for data annotation. To address this, techniques that enable model training directly on unlabeled data can be employed. Among these, constructing self-supervised attention mechanisms is a feasible approach, as it can learn transferable attention representations from unlabeled data. At the same time, integrating meta-learning and distillation strategies to develop few-shot attention models can enhance generalization capability in data-scarce scenarios. Furthermore, the emergence of large-scale pre-trained models (e.g., BioMedGPT) offers a paradigm shift [133]. Specifically, a research pathway can be outlined as follows: (1) self-supervised objectives are designed to learn anatomical priors; (2) meta-attention mechanisms adapt to new tasks with minimal labeled data; (3) knowledge distillation transfers attention patterns from large foundation models to compact clinical models. Experimental validation could involve benchmarking on standardized few-shot medical image classification tasks.

4.2.2 Domain-Adaptive and Generalizable Attention Mechanisms

To address the performance degradation of attention mechanisms caused by domain shifts in medical data, it is essential to develop domain-adaptive attention. This approach is expected to enable the model to automatically adjust attention weights as data distributions change, thereby improving transfer and generalization capabilities. In addition, integrating techniques such as normalization and adversarial learning can further enhance the consistency and stability of attention mechanisms when handling cross-domain features. A concrete implementation framework could involve: (1) domain-invariant attention learning through adversarial training; (2) test-time attention adaptation using batch statistics [134]; (3) federated attention learning across multiple medical centers while preserving data privacy. Quantitative evaluation can measure attention consistency across domains using metrics such as Attention Distribution Distance.

4.2.3 Dynamic Multi-Modal Fusion Attention

To effectively utilize information from various types of medical images and achieve high-quality multi-modal fusion, modality-adaptive cross-attention mechanisms can be designed to dynamically adjust feature contributions from different modalities. In addition, introducing modality gating can effectively address alignment issues between different image types and enable effective weighted fusion. Furthermore, exploring multimodal attention maps could help improve the interpretability of the multi-modal fusion process. This direction aligns with the trend of medical multi-modal pre-training. Large-scale models like Med-Aligner, which unify various medical imaging modalities and even integrate imaging with clinical or textual data, demonstrate the potential of unified multi-modal frameworks [135].

Future work can focus on developing modality-aware attention mechanisms that handle heterogeneous data types within unified frameworks such as BioMedGPT [133]. A specific research pathway involves:

(1) designing cross-modal attention layers with modality-specific projections; (2) implementing dynamic fusion gates based on modality reliability; (3) creating unified attention visualization tools for multi-modal diagnostics. Case studies could demonstrate improved diagnostic accuracy in complex cases requiring multi-modal integration [136,137].

4.2.4 Interpretable and Verifiable Attention Mechanisms

To enhance the medical interpretability of attention mechanisms, it is essential to establish semantically aligned interpretation frameworks. These frameworks can align attention weights with clinical features, such as anatomical structures and lesion areas. Specifically, integrating gradient-based visualization techniques (e.g., Grad-CAM) with attention mechanisms can create hybrid saliency maps that combine the strengths of both approaches. Additionally, developing attention consistency metrics can quantify the alignment between attention maps and expert annotations. In addition, introducing supervised attention and saliency map consistency constraints can improve the stability of the model's focus on relevant regions. To further increase healthcare professionals' trust in the decision-making process of attention mechanisms, developing interactive visualization systems could help clinicians better understand and accept AI-assisted diagnosis. Case studies can demonstrate how these interpretable attention mechanisms improve clinician confidence in AI-assisted diagnosis across different clinical specialties.

4.2.5 Efficient and Lightweight Attention Architectures

To enable model deployment on medical devices with limited computational resources, optimized attention mechanisms, such as linear or sparse attention, can be adopted to reduce computational load. In addition, hierarchical or shared attention modules can be designed to minimize parameter redundancy. The most effective approach involves developing hardware-friendly attention networks specifically tailored for mobile and clinical deployment, thereby achieving efficient implementation on medical equipment. This is particularly crucial for deploying large pre-trained models in clinical settings, where techniques like model compression, knowledge distillation, and dynamic inference will be key to harnessing the power of foundation models without prohibitive computational costs. A specific research pathway includes: (1) developing medical-specific efficient attention variants that exploit spatial sparsity in anatomical structures; (2) creating hardware-aware neural architecture search for attention models; (3) implementing progressive attention mechanisms that adapt computational cost based on image complexity.

4.2.6 Robust and Trustworthy Attention Models

By developing robust attention mechanisms, interference from common medical imaging artifacts, such as noise can be effectively suppressed, thereby enhancing model robustness. To further improve generalization and prevent over-reliance on specific features, uncertainty modeling techniques can be applied to constrain attention distributions within confidence intervals. In addition, incorporating adversarial training into attention frameworks will significantly strengthen the model's resistance to disturbances. Experimental validation can include stress testing attention models under various corruption types and severity levels, with quantitative metrics for attention stability and prediction reliability.

4.2.7 Integration with Large-Scale Models and Unified Medical AI Systems

Beyond the research directions mentioned above, the field is rapidly evolving towards large-scale, integrative AI systems. A prominent future direction lies in the evolution of attention mechanisms

within large-scale multi-modal pre-trained models for medicine. The goal is to construct unified vision-language-clinical foundation models like BioMedGPT, where advanced cross-modal attention serves as the fundamental mechanism for aligning and reasoning over heterogeneous data. Furthermore, generative AI models highlight the expanding role of attention in unified multi-modal generation and analysis, opening new avenues for data augmentation and synthetic data generation to combat scarcity.

Recent advances in large-scale medical foundation models demonstrate the feasibility of these approaches, though significant challenges remain in clinical deployment [133,135,136]. To realize this vision, specific implementation frameworks can focus on three key pillars:

(1) Federated and Privacy-Preserving Attention Learning: Developing specialized federated learning protocols for attention-based models is crucial. This involves designing methods to aggregate attention maps or their statistical distributions across multiple hospitals without sharing raw data. Techniques like secure multi-party computation or differential privacy could be applied to the attention weights themselves, ensuring that the model learns what to focus on from a global population while preserving patient confidentiality.

(2) Anatomy-Guided and Knowledge-Grounded Attention: Moving beyond purely data-driven attention, clinical knowledge can be explicitly incorporated into the attention mechanism. This can be achieved by initializing attention queries with embeddings of anatomical landmarks or by using segmentation masks of organs as a prior to constrain the spatial regions where attention can be allocated. For instance, when analyzing a chest X-ray, the model's attention can be softly guided towards lung fields, preventing it from being distracted by irrelevant regions and enhancing the biological plausibility of its focus.

(3) Attention as the Core of End-to-End Clinical Workflows: Future systems can position attention as the central, interpretable component that connects different stages of a clinical pipeline. For example, in a diagnostic workflow, the attention maps from a classification model (e.g., highlighting a potential nodule) can be directly fed into a segmentation model to precisely outline the lesion, and then used by a report generation model to justify the finding in natural language. This creates a transparent, "attention-aware" workflow where the reasoning process is traceable and verifiable by clinicians at every step.

The convergence of attention mechanisms with these strategic directions—federated learning for scalable and ethical data utilization, knowledge grounding for clinical trust, and workflow integration for practical utility—will define the next generation of medical AI. This integrated approach will transform attention from a mere performance-enhancing module into the foundational bedrock for building reliable, interpretable, and clinically admissible decision-support systems.

5 Conclusion

Attention mechanisms represent a significant technological breakthrough in medical image analysis, substantially enhancing models' ability to focus on key features and enabling efficient information fusion across multimodal data. They have markedly improved model performance and decision interpretability across multiple tasks—including classification, segmentation, reconstruction, and report generation—laying a solid foundation for intelligent medical imaging and precision-assisted diagnosis. Through continuous structural innovation, attention models have evolved from early single-space or single-channel focus to complex systems featuring multi-scale, cross-modal, and adaptive hybrid capabilities, powerfully expanding the application boundaries of medical image analysis.

Looking ahead, research on medical image attention mechanisms is expected to advance toward three critical objectives: efficiency, interpretability, and clinical integration. On one hand, efficient sparse attention and lightweight model designs hold promise for alleviating computational burdens imposed by high-resolution medical images, enhancing practical applicability. On the other hand, model interpretability

and decision transparency can be integrated with clinical knowledge to build intelligent systems trusted by physicians. Furthermore, addressing generalization capabilities in clinical settings, data privacy, and real-world validation will accelerate the translation of attention mechanisms into medical imaging practice. Moving forward, interdisciplinary collaboration and data resource sharing will become pivotal drivers for advancing the intelligent transformation of medical imaging.

Acknowledgement: None.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial, ornot-for-proft sectors.

Author Contributions: Xinjie Yao: writing—review & editing original draft. Junjie Zhu: review & editing. Tao Hong: review & editing. Dengyu Zhao: review & editing. Weikai Liu: review & editing. Guangsheng Xie: Supervision, Project administration. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This paper is predominantly a review that synthesizes existing methods and literature findings. This investigation utilized only data obtained from publicly accessible sources. These datasets are accessible via the sources listed in the References section of this paper.

Ethics Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

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

References

1. Müller H. Medical image retrieval: applications and resources. In: Proceedings of the 2020 International Conference on Multimedia Retrieval; 2020 Jun 8–11; Dublin, Ireland. p. 2–3. doi:10.1145/3372278.3390668.
2. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE. 2021;109(5):820–38. doi:10.1109/jproc.2021.3054390.
3. Wang J, Zhou P, Han X, Chen Y. Medical image super-resolution *via* diagnosis-guided attention. In: Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME); 2023 Jul 10–14; Brisbane, Australia. p. 462–7. doi:10.1109/icme55011.2023.00086.
4. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Cham, Switzerland: Springer International Publishing; 2015. p. 234–41. doi:10.1007/978-3-319-24574-4_28.
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42(13):60–88. doi:10.1016/j.media.2017.07.005.
6. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19(1):221–48. doi:10.1146/annurev-bioeng-071516-044442.
7. Çiçek Ö., Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 424–32. doi:10.1007/978-3-319-46723-8_49.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. doi:10.1038/nature21056.
9. Wen Y, Chen L, Chen H, Tang X, Deng Y, Chen Y, et al. Non-local attention learning for medical image classification. In: Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME); 2021 Jul 5–9; Shenzhen, China. p. 1–6. doi:10.1109/icme51207.2021.9428267.

10. Pal D, Meena T, Mahapatra D, Roy S. AW-net: a novel fully connected attention-based medical image segmentation model. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2023 Oct 2–6; Paris, France. p. 2532–41. doi:10.1109/iccvw60793.2023.00267.
11. Fontanella A, Antoniou A, Li W, Wardlaw J, Mair G, Trucco E, et al. ACAT: adversarial counterfactual attention for classification and detection in medical imaging. arXiv:230315421. 2023. doi:10.48550/arXiv.2303.15421.
12. Li D, Yang X, Chen S, Deng L, Lan Q, Huang S, et al. TSMR-Net: a two-stage multimodal medical image registration method *via* pseudo-image generation and deformable registration. Pattern Recognit Lett. 2025;197(8):359–67. doi:10.1016/j.patrec.2025.09.006.
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. arXiv:1706.03762. 2017. doi:10.48550/arXiv.1706.03762.
14. Dosovitskiy A. An image is worth 16×16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020. doi:10.48550/arXiv.2010.11929.
15. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22. doi:10.1109/iccv48922.2021.00986.
16. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal. 2024;97(2):103280. doi:10.1016/j.media.2024.103280.
17. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision—ECCV 2022 Workshops. Cham, Switzerland: Springer Nature; 2023. p. 205–18. doi:10.1007/978-3-031-25066-8_9.
18. Wu J, Xu M. One-prompt to segment all medical images. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 11302–12. doi:10.1109/cvpr52733.2024.01074.
19. Wang Z, Wang J, Song H, Feng J, Duan H. Multi-modal medical image fusion *via* 3D manifold fitting and dual-domain cross-attention. In: Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5. doi:10.1109/icassp49660.2025.10888400.
20. Shiraishi T, Miwa D, Katsuoka T, Duy VNL, Taji K, Takeuchi I. Statistical test for attention map in vision transformer. arXiv:2401.08169. 2024. doi:10.48550/arXiv.2401.08169.
21. Qiu X, Liang D, Luo G, Li X, Wang W, Wang K, et al. MeMGB-Diff: memory-efficient multivariate Gaussian bias diffusion model for 3D bias field correction. Med Image Anal. 2025;102(4):103560. doi:10.1016/j.media.2025.103560.
22. Chen Y, Liu J, Zuo Z, Jiang P, Jin Y, Wu G. Classifying pathological images based on multi-instance learning and end-to-end attention pooling. In: Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. Greece: Rhodes Island. p. 1–5. doi:10.1109/icassp49357.2023.10094946.
23. Zhao Y, Zhou X, Guo H, Guo Q, Zuo Y, Song S, et al. Attention in attention for PET-CT modality consensus lung tumor segmentation. In: Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME); 2024 Jul 15–19; Niagara Falls, ON, Canada. p. 1–7. doi:10.1109/icme57554.2024.10687909.
24. Zhu L, Wu X, Wang C, Wang H. SAM adaptation with refocused attention and diverse prompts for medical image segmentation. In: Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5. doi:10.1109/icassp49660.2025.10889427.
25. Huang Z, Chen B. Unsupervised multi-modal medical image registration via query-selected attention and decoupled contrastive learning. In: Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME); 2024 Jul 15–19; Niagara Falls, ON, Canada. p. 1–6. doi:10.1109/icme57554.2024.10688363.
26. Song X, Zhang X, Ji J, Liu Y, Wei P. Cross-modal contrastive attention model for medical report generation. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022 Oct 12–17; Gyeongju, Republic of Korea. p. 2388–97.

27. Yan B, Pei M. Clinical-BERT: vision-language pre-training for radiograph diagnosis and reports generation. Proc AAAI Conf Artif Intell. 2022;36(3):2982–90. doi:10.1609/aaai.v36i3.20204.
28. Graves A, Wayne G, Danihelka I. Neural turing machines. arXiv:14105401. 2014.
29. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078. 2014.
30. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. 2014. doi:10.48550/arXiv.1409.0473.
31. Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, et al. Attention mechanisms in computer vision: a survey. Comp Visual Med. 2022;8(3):331–68. doi:10.1007/s41095-022-0271-y.
32. Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-alone self-attention in vision models. In: Advances in neural information processing systems. Vol. 32. Red Hook, NY, USA: Curran Associates, Inc.; 2019. p. 68–80.
33. Clark K, Khandelwal U, Levy O, Manning CD. What does Bert look at? An analysis of Bert's attention. arXiv:1906.04341. 2019.
34. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 782–91. doi:10.1109/cvpr46437.2021.00084.
35. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA. p. 7132–41. doi:10.1109/cvpr.2018.00745.
36. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11534–42. doi:10.1109/cvpr42600.2020.01155.
37. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Computer Vision—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
38. Li X, Wang W, Hu X, Yang J. Selective kernel networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 510–9. doi:10.1109/cvpr.2019.00060.
39. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans Med Imaging. 2019;38(8):1788–800. doi:10.1109/tmi.2019.2897538.
40. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2021. Cham, Switzerland: Springer International Publishing; 2021. p. 36–46. doi:10.1007/978-3-030-87193-2_4.
41. Zafari-Ghadim Y, Rashed EA, Mohamed A, Mabrok M. Transformers-based architectures for stroke segmentation: a review. Artif Intell Rev. 2024;57(11):307. doi:10.1007/s10462-024-10900-5.
42. Li J, Xu Q, He X, Liu Z, Zhang D, Wang R, et al. CFFormer: cross CNN-Transformer channel attention and spatial feature fusion for improved segmentation of heterogeneous medical images. Expert Syst Appl. 2026;295(1):128835. doi:10.1016/j.eswa.2025.128835.
43. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? arXiv:2102.05095. 2021.
44. Sukegawa S, Yoshii K, Hara T, Tanaka F, Yamashita K, Kagaya T, et al. Is attention branch network effective in classifying dental implants from panoramic radiograph images by deep learning? PLoS One. 2022;17(7):e0269016. doi:10.1371/journal.pone.0269016.
45. Zhang C, Wang L, Wei G, Kong Z, Qiu M. A dual-branch and dual attention transformer and CNN hybrid network for ultrasound image segmentation. Front Physiol. 2024;15:1432987. doi:10.3389/fphys.2024.1432987.
46. Cao Y, Zhou W, Zang M, An D, Feng Y, Yu B. MBANet: a 3D convolutional neural network with multi-branch attention for brain tumor segmentation from MRI images. Biomed Signal Process Control. 2023;80(2):104296. doi:10.1016/j.bspc.2022.104296.

47. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, et al. Attention gated networks: learning to leverage salient regions in medical images. Med Image Anal. 2019;53(7639):197–207. doi:10.1016/j.media.2019.01.012.
48. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 6–11; Lille, France. p. 2048–57.
49. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, et al. Residual attention network for image classification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 3156–64. doi:10.1109/cvpr.2017.683.
50. Takahashi S, Sakaguchi Y, Kouno N, Takasawa K, Ishizu K, Akagi Y, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J Med Syst. 2024;48(1):84. doi:10.1007/s10916-024-02105-8.
51. Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv:1508.04025. 2015.
52. Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. arXiv:1803.02155. 2018.
53. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv:2004.05150. 2020.
54. Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, et al. Image transformer. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 4055–64.
55. Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, et al. CvT: introducing convolutions to vision transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17. Montreal, QC, Canada. p. 22–31. doi:10.1109/iccv48922.2021.00009.
56. Srinivas A, Lin TY, Parmar N, Shlens J, Abbeel P, Vaswani A. Bottleneck transformers for visual recognition. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25. Nashville, TN, USA. p. 16519–29. doi:10.1109/cvpr46437.2021.01625
57. Chen Y, Dai X, Chen D, Liu M, Dong X, Yuan L, et al. Mobile-former: bridging MobileNet and transformer. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24. New Orleans, LA, USA. p. 5270–9. doi:10.1109/cvpr52688.2022.00520.
58. Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jegou H, et al. LeViT: a vision transformer in ConvNet's clothing for faster inference. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17. Montreal, QC, Canada. p. 12259–69. doi:10.1109/iccv48922.2021.01204
59. Khan A, Rauf Z, Sohail A, Khan AR, Asif H, Asif A, et al. A survey of the vision transformers and their CNN-transformer based variants. Artif Intell Rev. 2023;56(3):2917–70. doi:10.1007/s10462-023-10595-0.
60. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Jun 12–17. San Diego, CA, USA. p. 1480–9. doi:10.18653/v1/n16-1174
61. Lin Z, Feng M, Santos CNd, Yu M, Xiang B, Zhou B, et al. A structured self-attentive sentence embedding. arXiv:1703.03130. 2017.
62. Yuan Y, Fu R, Huang L, Lin W, Zhang C, Chen X, et al. Hrformer: high-resolution vision transformer for dense predict. Adv Neural Inf Process Syst. 2021;34:7281–93.
63. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA. p. 7794–803. doi:10.1109/cvpr.2018.00813
64. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. arXiv:1804.03999. 2018. doi:10.48550/arXiv.1804.03999.
65. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306. 2021. doi:10.48550/arXiv.2102.04306.
66. Chen S, Ma K, Zheng Y. Med3d: transfer learning for 3D medical image analysis. arXiv:1904.00625. 2019.

67. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 2127–36.
68. Zhang Y, Ying MTC, Yang L, Ahuja AT, Chen DZ. Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images. In: Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2016 Dec 15–18; Shenzhen, China. p. 443–8. doi:10.1109/bibm.2016.7822557
69. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: transformers for 3D medical image segmentation. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA. p. 574–84. doi:10.1109/wacv51458.2022.00181
70. Wang W, Chen C, Ding M, Yu H, Zha S, TransBTS Li J. Multimodal brain tumor segmentation using transformer. In: Medical image computing and computer assisted intervention—MICCAI 2021. Cham, Switzerland: Springer International Publishing; 2021. p. 109–19. doi:10.1007/978-3-030-87193-2_11.
71. Song Y, Lu Y, Chen L, Luo Y. Hierarchical multi-scale enhanced transformer for medical image segmentation. IEEE J Biomed Health Inform. 2025;29(12):8917–27. doi:10.1109/jbhi.2024.3515477.
72. Valindria VV, Pawlowski N, Rajchl M, Lavdas I, Aboagye EO, Rockall AG, et al. Multi-modal learning from unpaired images: application to multi-organ segmentation in CT and MRI. In: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); 2018 Mar 12–15; Lake Tahoe, NV, USA. p. 547–56. doi:10.1109/wacv.2018.00066
73. Chartsias A, Joyce T, Giuffrida MV, Tsaftaris SA. Multimodal MR synthesis via modality-invariant latent representation. IEEE Trans Med Imaging. 2018;37(3):803–14. doi:10.1109/tmi.2017.2764326.
74. Dou Q, Ouyang C, Chen C, Chen H, Heng PA. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. arXiv:180410916. 2018.
75. Kamnitsas K, Baumgartner C, Ledig C, Newcombe V, Simpson J, Kane A, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging. Cham, Switzerland: Springer International Publishing; 2017. p. 597–609. doi:10.1007/978-3-319-59050-9_47.
76. Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol. 2021;66(4):04TR01. doi:10.1088/1361-6560/abcd17.
77. Baumgartner CF, Koch LM, Tezcan KC, Ang JX, Konukoglu E. Visual feature attribution using Wasserstein GANs. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 8309–19. doi:10.1109/cvpr.2018.00867.
78. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Für Med Phys. 2019;29(2):102–27. doi:10.1016/j.zemedi.2018.11.002.
79. Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. TransAttUnet: multi-level attention-guided U-Net with transformer for medical image segmentation. IEEE Trans Emerg Top Comput Intell. 2024;8(1):55–68. doi:10.1109/tetci.2023.3309626.
80. Lou A, Guan S, Loew M. CaraNet: context axial reverse attention network for segmentation of small medical objects. Med Imaging 2022: Image Process. 2022;12032:81–92.
81. Dai W, Liu R, Wu Z, Wu T, Wang M, Zhou J, et al. Exploiting scale-variant attention for segmenting small medical objects. arXiv:240707720. 2024. doi:10.48550/arxiv.2407.07720.
82. Zhao L, Wang T, Chen Y, Zhang X, Tang H, Lin F, et al. A novel framework for segmentation of small targets in medical images. Sci Rep. 2025;15(1):9924. doi:10.1038/s41598-025-94437-9.
83. Shao H, Zeng Q, Hou Q, Yang J. MCANet: medical image segmentation with multi-scale cross-axis attention. Mach Intell Res. 2025;22(3):437–51. doi:10.1007/s11633-025-1552-6.
84. Zheng F, Chen X, Liu W, Li H, Lei Y, He J, et al. SMAFormer: synergistic multi-attention transformer for medical image segmentation. In: Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2024 Dec 3–6; Lisbon, Portugal. p. 4048–53. doi:10.1109/bibm62325.2024.10822736
85. Roy AG, Navab N, Wachinger C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2018. Cham, Switzerland: Springer International Publishing; 2018. p. 421–9. doi:10.1007/978-3-030-00928-1_48.

86. Huo X, Sun G, Tian S, Wang Y, Yu L, Long J, et al. HiFuse: hierarchical multi-scale feature fusion network for medical image classification. Biomed Signal Process Control. 2024;87(7660):105534. doi:10.1016/j.bspc.2023.105534.
87. Manzari ON, Ahmadabadi H, Kashiani H, Shokouhi SB, Ayatollahi A. MedViT: a robust vision transformer for generalized medical image classification. Comput Biol Med. 2023;157(9):106791. doi:10.1016/j.compbiomed.2023.106791.
88. Nejati Manzari O, Asgariandehkordi H, Koleilat T, Xiao Y, Rivaz H. Medical image classification with KAN-integrated transformers and dilated neighborhood attention. Appl Soft Comput. 2026;186(11):114045. doi:10.1016/j.asoc.2025.114045.
89. Gu P, Yao T, He M, Duan F, Liu F, Peng R, et al. SwinECAT: a transformer-based fundus disease classification model with shifted window attention and efficient channel attention. arXiv:2507.21922. 2025. doi:10.48550/arXiv.2507.21922.
90. Wu X, Feng Y, Xu H, Lin Z, Chen T, Li S, et al. CTransCNN: combining transformer and CNN in multilabel medical image classification. Knowl Based Syst. 2023;281(2):111030. doi:10.1016/j.knosys.2023.111030.
91. Zhao Z, Bai H, Zhang J, Zhang Y, Xu S, Lin Z, et al. CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24. Vancouver, BC, Canada. p. 5906–16. doi:10.1109/cvpr52729.2023.00572
92. Wang X, Feng Y, Wang S, Wang D, Cheng TCE. An explainable lesion detection transformer model for medical imaging diagnosis decision support: design science research. Decis Support Syst. 2025;196:114492. doi:10.1016/j.dss.2025.114492.
93. Zhao Y, Zhou Z, Qi L, Xue H. Precision in pathology: PMA-DETR elevates tumor lesion detection. In: Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5. doi:10.1109/icassp49660.2025.10887994.
94. Russo C, Bria A, Marrocco C. GravityNet for end-to-end small lesion detection. Artif Intell Med. 2024;150(2):102842. doi:10.1016/j.artmed.2024.102842.
95. Sobek J, Medina Inojosa JR, Medina Inojosa BJ, Rassoulinejad-Mousavi SM, Conte GM, Lopez-Jimenez F, et al. MedYOLO: a medical image object detection framework. J Digit Imaging Inform Med. 2024;37(6):3208–16. doi:10.1007/s10278-024-01138-2.
96. Li YX, Tang H, Wang W, Zhang XF, Qu H. Dual attention network for unsupervised medical image registration based on VoxelMorph. Sci Rep. 2022;12(1):16250. doi:10.1038/s41598-022-20589-7.
97. Chen J, Frey EC, He Y, Segars WP, Li Y, Du Y. TransMorph: transformer for unsupervised medical image registration. Med Image Anal. 2022;82(5):102615. doi:10.1016/j.media.2022.102615.
98. Zhuo Y, Shen Y. DiffuseReg: denoising diffusion model for obtaining deformation fields in unsupervised deformable image registration. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 597–607. doi:10.1007/978-3-031-72069-7_56.
99. Yang G, Yu S, Dong H, Slabaugh G, Dragotti PL, Ye X, et al. DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Trans Med Imaging. 2018;37(6):1310–21. doi:10.1109/tmi.2017.2785879.
100. Emami H, Aliabadi MM, Dong M, Chinnam RB. SPA-GAN: spatial attention GAN for image-to-image translation. IEEE Trans Multimedia. 2021;23:391–401. doi:10.1109/tmm.2020.2975961.
101. Upadhyay U, Chen Y, Hepp T, Gatidis S, Akata Z. Uncertainty-guided progressive GANs for medical image translation. In: Medical image computing and computer assisted intervention—MICCAI 2021. Cham, Switzerland: Springer International Publishing; 2021. p. 614–24. doi:10.1007/978-3-030-87199-4_58.
102. Huang J, Fang Y, Wu Y, Wu H, Gao Z, Li Y, et al. Swin transformer for fast MRI. Neurocomputing. 2022;493(5028):281–304. doi:10.1016/j.neucom.2022.04.051.
103. Fei B, Li Y, Yang W, Gao H, Xu J, Ma L, et al. A diffusion model for universal medical image enhancement. Commun Med. 2025;5(1):294. doi:10.1038/s43856-025-00998-1.

104. Wang H, Liu Z, Sun K, Wang X, Shen D, Cui Z. 3D meddiffusion: a 3D medical diffusion model for controllable and high-quality medical image generation. arXiv:2412.13059. 2024. doi:10.48550/arXiv.2412.13059.
105. Dai Y, Gao Y, Liu F. TransMed: transformers advance multi-modal medical image classification. Diagnostics. 2021;11(8):1384. doi:10.3390/diagnostics11081384.
106. Tang W, He F, Liu Y, Duan Y. MATR: multimodal medical image fusion *via* multiscale adaptive transformer. IEEE Trans Image Process. 2022;31(12):5134–49. doi:10.1109/tip.2022.3193288.
107. He D, Li W, Wang G, Huang Y, Liu S. DM-FNet: unified multimodal medical image fusion via diffusion process-trained encoder-decoder. IEEE Trans Multimedia 2025;27:9415–28. doi:10.1109/tmm.2025.3613156.
108. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11. Abu Dhabi, United Arab Emirates. p. 3876. doi:10.18653/v1/2022.emnlp-main.256
109. Vardhanabhuti V, Wang F, Wang S, Yu L, Zhou Y. Multi-granularity cross-modal alignment for generalized medical visual representation learning. Adv Neural Inf Process Syst. 2022;35:33536–49. doi:10.52202/068431-2430.
110. Bannur S, Hyland S, Liu Q, Pérez-García F, Ilse M, Castro DC, et al. Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 15016–27. doi:10.1109/cvpr52729.2023.01442
111. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: a survey. IET Image Process. 2022;16(5):1243–67. doi:10.1049/ipr2.12419.
112. Zhao Z, Chen K, Yamane S. CBAM-Unet++: easier to find the target with the attention module "CBAM". In: Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE); 2021 Oct 12–15; Kyoto, Japan. p. 655–7. doi:10.1109/gcce53005.2021.9622008
113. Heidari M, Kazerouni A, Soltany M, Azad R, Aghdam EK, Cohen-Adad J, et al. HiFormer: hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 6202–12. doi:10.1109/wacv56688.2023.00614
114. Zhou HY, Guo J, Zhang Y, Han X, Yu L, Wang L, et al. nnFormer: volumetric medical image segmentation *via* a 3D transformer. IEEE Trans Image Process. 2023;32:4036–45. doi:10.1109/tip.2023.3293771.
115. Landman B, Warfield S. MICCAI 2012 workshop on multi-atlas labelin. Nice, France: CreateSpace Independent Publishing Platform; 2012.
116. Chowdary GJ, Yin Z. Med-former: a transformer based architecture for medical image classification. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 448–57. doi:10.1007/978-3-031-72120-5_42.
117. Li K, Qian Z, Han Y, Chang EI, Wei B, Lai M, et al. Weakly supervised histopathology image segmentation with self-attention. Med Image Anal. 2023;86(1–2):102791. doi:10.1016/j.media.2023.102791.
118. Luo X, Hu M, Song T, Wang G, Zhang S. Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In: Proceedings of the 5th International Conference on Medical Imaging with Deep Learning; 2022 Jul 6–8; Zurich, Switzerland. p. 820–33.
119. Yin S, Li H, Teng L, Ali Laghari A, Almadhor A, Gregus M, et al. Brain CT image classification based on mask RCNN and attention mechanism. Sci Rep. 2024;14(1):29300. doi:10.1038/s41598-024-78566-1.
120. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/iccv.2017.324
121. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993–2024. doi:10.1109/tmi.2014.2377694.
122. Armato IIISG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, 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(2):915–31.

123. Shi J, He Y, Kong Y, Coatrieux JL, Shu H, Yang G, et al. XMorpher: full transformer for deformable medical image registration via cross attention. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2022. Cham, Switzerland: Springer Nature; 2022. p. 217–26. doi:10.1007/978-3-031-16446-0_21.
124. Chen J, Liu Y, He Y, Du Y. Deformable cross-attention transformer for medical image registration. In: Machine Learning in Medical Imaging. Cham, Switzerland: Springer Nature; 2023. p. 115–25. doi:10.1007/978-3-031-45673-2_12.
125. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. MedKLIP: medical knowledge enhanced language-image pre-training for X-ray diagnosis. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 21372–83. doi:10.1109/iccv51070.2023.01954
126. Kim B, Kim DH, Park SH, Kim J, Lee JG, Ye JC. CycleMorph: cycle consistent unsupervised deformable image registration. Med Image Anal. 2021;71(1):102036. doi:10.1016/j.media.2021.102036.
127. Song W, Zeng X, Abdelmoniem AM, Zhang H, Gao M. Cross-modality interaction network for medical image fusion. IEEE Trans Consumer Electron. 2025;71(1):1385–92. doi:10.1109/tce.2024.3412879.
128. Kim B, Ye JC. Diffusion deformable model for 4D temporal medical image generation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2022. Cham, Switzerland: Springer Nature; 2022. p. 539–48. doi:10.1007/978-3-031-16431-6_51.
129. Ma J, Zhu Y, You C, Wang B. Pre-trained diffusion models for plug-and-play medical image enhancement. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2023. Cham, Switzerland: Springer Nature; 2023. p. 3–13. doi:10.1007/978-3-031-43898-1_1.
130. Zhou X, Song Q, Nie J, Feng Y, Liu H, Liang F, et al. Hybrid cross-modality fusion network for medical image segmentation with contrastive learning. Eng Appl Artif Intell. 2025;144(6):110073. doi:10.1016/j.engappai.2025.110073.
131. Su L, Ma X, Zhu X, Niu C, Lei Z, Zhou JZ. Can we get rid of handcrafted feature extractors? SparseViT: nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. Proc AAAI Conf Artif Intell. 2025;39(7):7024–32. doi:10.1609/aaai.v39i7.32754.
132. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 618–26. doi:10.1109/iccv.2017.74
133. Luo Y, Zhang J, Fan S, Yang K, Hong M, Wu Y, et al. Biomedgpt: an open multimodal large language model for biomedicine. IEEE J Biomed Health Inform 2024;30(2):981–992. doi:10.1109/JBHI.2024.3505955.
134. Wang Z, Lu Z, Wang T, Yang Z, Yu H, Wang Z, et al. Test-time adaptation via orthogonal meta-learning for medical imaging. IEEE Trans Radiat Plasma Med Sci. 2024;9(2):215–27. doi:10.1109/TRPMS.2024.3462542.
135. Meng X, Ji JM, Yan X, Dai JT, Chen BY, Wang G, et al. Med-aligner empowers LLM medical applications for complex medical scenarios. Innovation. 2025;6(11):101002. doi:10.1016/j.xinn.2025.101002.
136. Zhan C, Lin Y, Wang G, Wang H, Wu J. MedM2G: unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 11502–12. doi:10.1109/cvpr52733.2024.01093
137. Yao W, Lyu Z, Mahmud M, Zhong N, Lei B, Wang S. CATD: unified representation learning for EEG-to-fMRI cross-modal generation. IEEE Trans Med Imaging. 2025;44(7):2757–67. doi:10.1109/tmi.2025.3550206.