IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

VISION TRANSFORMERS IN MEDICAL IMAGING: A COMPREHENSIVE RESEARCH PAPER ON ARCHITECTURE, COMPARISONS, AND APPLICATIONS

[1] Sneha Kashyap, [2] Dr. Arvind Kumar Pandey

[1]Research Scholar, Dept. of Computer Science, ARKA JAIN University, Jharkhand, India

[2]Dean, School of Engineering & IT, ARKA JAIN University, Jharkhand, India

Abstract: The Transformer architecture, originally developed for Natural Language Processing, has emerged as a prominent new approach in computer vision. This review concentrates on Vision Transformers (ViTs) and their application in medical imaging. We discuss the fundamental Transformer components, including self-attention, multi-head attention, and encoder-decoder structures, then describe how images are divided into patch-based sequences for transformer processing. The review highlights key elements of ViT architecture, such as patch embedding, positional encoding, encoder design, and feed-forward layers, and compares ViTs to Convolutional Neural Networks (CNNs) regarding feature extraction, local versus global context understanding, efficiency, and benchmark results. We also examine hybrid models that combine CNNs' emphasis on local features with the broad modeling capacity of self-attention. Additionally, we review notable ViT models like the Swin Transformer, DeiT, and those tailored for medical tasks like TransUNet and UNETR. Our findings indicate that, although CNNs remain effective with limited data, ViTs and hybrid models generally outperform pure CNNs in large-scale, resource-intensive medical imaging tasks. These outcomes highlight the increasing importance of Vision Transformers in medical applications such as tumor detection, organ segmentation, and disease classification.

Index Terms - Vision Transformers, Medical Imaging, Self-Attention, Convolutional Neural Networks, Swin Transformer, DeiT, Patch Embedding, Deep Learning.

I. INTRODUCTION

Deep learning has fundamentally changed computer vision. For most of the 2010s, the model families were dominated by Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which were particularly good at sequence modeling and local feature extraction, respectively [7, 15]. However, CNNs are constrained by local receptive fields and struggle to simulate long-range spatial relationships, whereas RNNs suffer from vanishing gradients and lack of parallelization. In the 2017 study "Attention Is All You Need" by Vaswani et al. at Google Brain, recurrence was entirely replaced by the Transformer, a parallelizable self-attention mechanism that captures global dependencies in a single step [1].

Models that can both understand the global anatomical context and detect minute local anomalies are necessary for medical imaging, which includes MRI, CT, and X-ray modalities. These two characteristics make medical imaging a perfect test-bed for comparing CNNs (strong local priors) with ViTs (strong global modeling) [6, 10]. Every aspect of Vision Transformers, from their architectural foundations to their medicinal uses, is covered in this study's methodical analysis.

1.1 Introduction to Transformers

In the fast-moving domain of Artificial Intelligence, few developments have been as influential as the Transformer architecture [1]. The 2020s have truly become the Transformer era, driving innovations such as ChatGPT, Claude, Midjourney, and most major AI breakthroughs of the decade. The development of Transformer models originates from Google Brain's 2017 paper "Attention Is All You Need" [1]. Before this innovation, sequence processing depended on RNNs and LSTMs, which processed data sequentially and faced two major problems: (1) Vanishing Gradients, causing the model to "forget" earlier parts of long sequences; and (2) Limited Parallelization, as waiting for each step hindered efficient use of modern GPUs. The Transformer changed the game by removing recurrence and processing entire sequences at once through Attention. This approach allows the model to identify the most relevant parts of the input relative to each other, regardless of their distance [1].

1.2 Transformer Architecture: Encoder–Decoder Structure

The Transformer architecture is built on an Encoder-Decoder framework [1]. Many downstream models depend on just one part: GPT uses only the decoder, while BERT relies solely on the encoder. Vision Transformers primarily utilize the encoder.

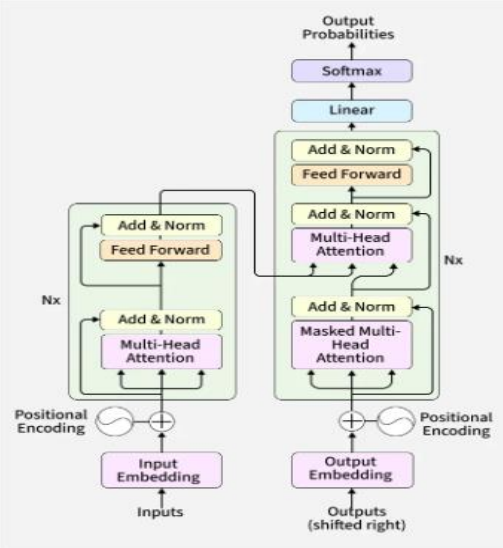

Figure 1: Transformer Architecture

The Encoder converts the input sequence into a detailed contextual representation. First, tokens are processed through an Input Embedding layer, followed by adding Positional Encoding, since the Transformer does not inherently understand sequence order. These embedded tokens then pass through Nx stacked Encoder blocks, each containing: (1) Multi-Head Self-Attention, allowing tokens to learn relationships with all others for context; (2) Add & Norm, involving residual connections and Layer Normalization to improve gradient stability; (3) a Position-wise Feed Forward network, which is a two-layer MLP; and (4) another Add & Norm [1].

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

The Decoder produces output tokens sequentially based on the encoder's representation. It incorporates Masked Multi-Head Attention, which uses a causal mask to prevent attention to future tokens, and Encoder-Decoder Cross-Attention, where the Query (Q) originates from the decoder's previous layer, while the Key (K) and Value (V) are derived from the encoder's output. This design allows the decoder to focus on relevant parts of the input during each token generation. The final decoder layer applies a Linear projection followed by a Softmax to generate output probabilities over the vocabulary [1].

Building Blocks Summary

A. Multi-Head Attention (MHA): Information is processed concurrently by the "brain" from multiple perspectives. One head may focus on semantics, while another may focus on syntax [1].

B. Feed-Forward Networks (FFN): This filter refines attention outputs to convert unprocessed attended data into meaningful representations [1].

C. Layer Norm & Residual Connections: Over 50–100 layers, the "stabilizers" guard against signal loss or distortion [1].

1.3 Attention Mechanisms: Scaled Dot-Product & The Cocktail Party Effect

The "Cocktail Party Effect" is a useful metaphor: in a crowded room, your brain ignores background noise and focuses just on the person you are speaking to. Transformers employ three vectors to do this mathematically: Query (Q), Key (K), and Value (V) [1]. Think about a library: Q is what you're looking for, K is the label on each book's spine, and V is the information contained within. The model determines a match score between Q and all Ks by taking into account the corresponding V. In mathematical terms:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Because the scaling factor 〖√d〗_k prevents dot products from being too large, learning gradients continue to be responsive [1]. Self-attention links locations inside a single sequence, like "it" and "animal" in a sentence. The decoder connects the locations of two sequences via cross-attention. A translator can refer to the original sentence, for example [1].

1.4 Applications of Transformers in Computer Vision

In 2020, ViT questioned the notion that CNNs are exclusively for images while Transformers are exclusively for text [7, 12]. An image is processed by ViT in the same way as a sentence: (1) it is divided into 16x16 patches; (2) these patches are flattened into vectors; (3) positional embeddings are added; and (4) a Transformer encoder is applied. The following are some of the main applications of computer vision: video analysis (video as a series of images, perfect for Transformers); object detection using DETR (treating detection as set prediction) [12]; image segmentation using Swin Transformer (using hierarchical windows) [12]; and generative models like diffusion models like Stable Diffusion. Because transformers are scalable, their attention method can be applied uniformly to both text and pixels, and more data and compute always improves speed, they have become dominant.

II. Architecture of Vision Transformers (ViT)

By applying the Transformer—which was first created for natural language processing—to image analysis, Vision Transformers (ViT) mark a substantial breakthrough [7, 12]. CNNs are limited in their ability to describe long-range dependencies and global context across a picture; instead, they rely on convolutions to capture spatial hierarchies and local characteristics. ViT tackles this issue by treating images as patch sequences and immediately applying self-attention to image data [7, 12]. ViT, which was first presented by Google Research in 2020 in "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," showed cutting-edge results for large-scale image categorization [7]. ViT has drawn interest in medical imaging because it can model intricate spatial relationships in MRI, CT, and X-ray images, making it possible to identify tumors, segment organs, classify diseases, and identify anomalies [6, 12].

2.1 Overview: Key Components

The ViT pipeline consists of Image Patch Extraction, Patch Embedding Layer, Positional Encoding, Transformer Encoder Blocks, Multi-Head Self-Attention, Feed-Forward Networks, and Classification Head. Unlike CNNs, which use local receptive fields, ViT uses self-attention to evaluate all patch interactions simultaneously, collecting global contextual information and long-range interdependence [7, 12]. A distinct CLS token is created by combining data from all patches and appending it to the sequence; this final representation is used for classification [7].

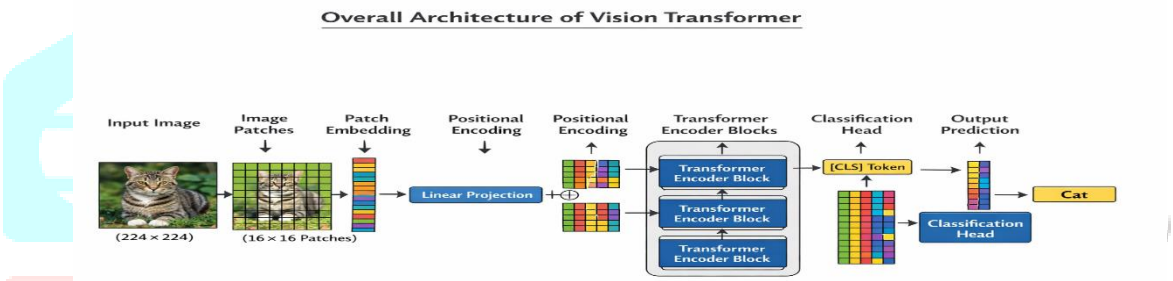

Figure 2 : Architecture of Vision Transformer

2.2 Image-to-Sequence Transformation

Given an image H × W × C, patches of size P × P are extracted. The number of patches N = (H × W) / P². For a 224×224×3 image with patch size 16×16:

$$\mathbf{N = \frac{224*224}{16*16} = 196}$$

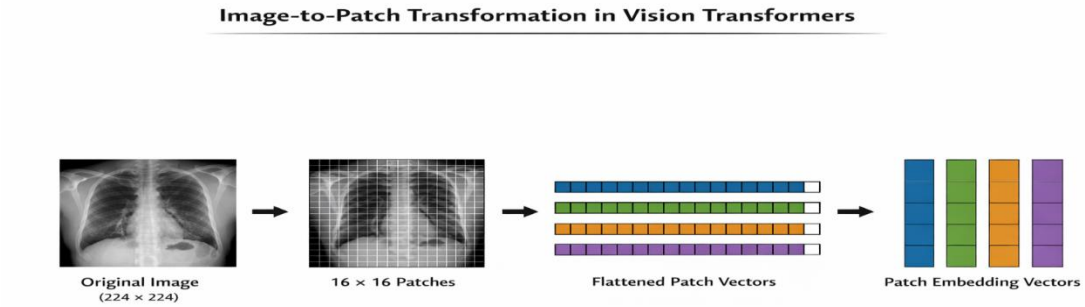

Figure 3: Image-to-patch in vision transformers

Each patch is flattened into a vector of length $P^2 \bullet C = 16 \times 16 \times 3 = 768$. A linear projection layer embeds each into D-dimensional space. The final input sequence , with the CLS token prepended , is [CLS,x1,x2,x3,....,xn] [7].

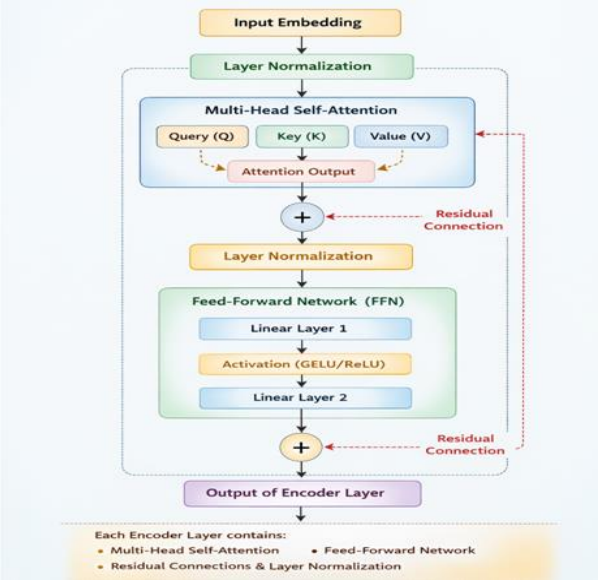

Figure 4: Internal structure of a transformer encoder layer used in Vision Transformers

2.3 ENCODER, PATCH EMBEDDING, AND POSITIONAL ENCODING

THE TRANSFORMER IS COMPOSED OF L STACKED IDENTICAL BLOCKS (OFTEN L=12 OR 24). LAYER NORM → MULTI-HEAD SELF-ATTENTION → RESIDUAL → LAYER NORM → FFN → RESIDUAL IS PRESENT IN EVERY BLOCK. REMAINING LINKS ENSURE CONSISTENT GRADIENT FLOW IN VERY DEEP NETWORKS [1, 7].

Patch Embedding: The linear projection of each flattened patch xi is zi = xi •E, where E is a learnable projection matrix. This is similar to word embedding in NLP. Patch size selection is crucial; smaller patches need more processing resources but capture finer information [7].Positional Encoding: Because Transformers handle all tokens at once without sequential ordering, positional information must be manually supplied [1, 7]. There are two methods: (1) Fixed, deterministic sine/cosine functions generate distinct positional vectors; (2) Learnable, positional embeddings are trainable parameters optimized during training. Without this, the model sees patches as an unordered set and loses all spatial organization required for medical picture interpretation [6, 7].

$$z_i = x_i E$$

2.4 Self-Attention, Multi-Head Attention, and Feed-Forward Layers

By deciding how vigorously each patch should pay attention to every other patch, self-attention captures global dependencies throughout the entire image [1, 7]. Three vectors, Q, K, and V, are produced by linearly projecting the input embeddings:

$$\mathbf{Q = XW_Q,\ K = XW_K,\ V = XW_V}$$

$$\mathbf{Attention(Q,K,V) = softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathbf{V}$$

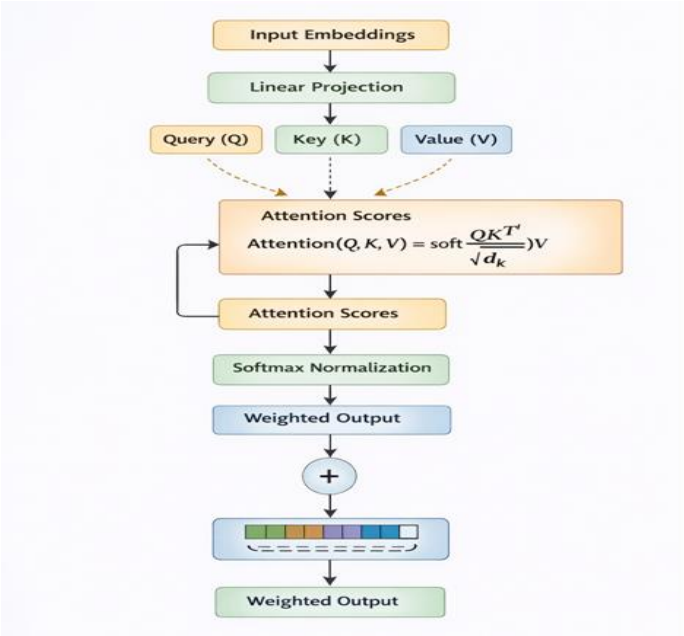

Figure 5: Self-attention mechanism showing Query, Key, and Value computation and attention score calculation

Multi-Head Attention: Rather than using a single attention function, parallel heads learn various associations. Some heads store local patterns, whereas others capture global context [1, 7]:

Multihead (Q,K,V)= Concat(head$_1$,head$_2$,head$_3$.....head$_h$)W^0

Where, head$_i$ = Attention(QW_i^Q,KW_i^K,VW_i^V)

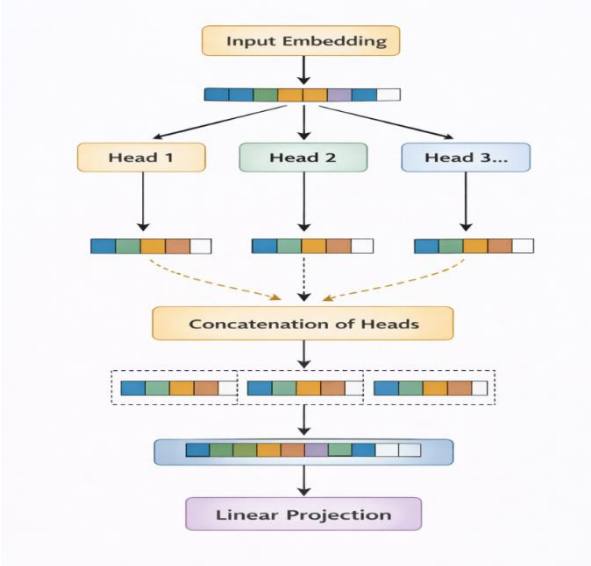

Figure 6: Multi-head attention mechanism where multiple attention heads learn different feature relationships.

Multi-head attention enhances feature extraction, spatial connection modeling, and robustness in complex images, making it highly valuable when examining multiple anatomical structures simultaneously in medical imaging [6, 12].

Feed-Forward Layers: Each encoder block has a position-wise FFN using GELU or ReLU activation: **FFN(x) = max(0,xW$_1$+b$_1$)W$_2$+b$_2$**, By enhancing representational strength and honing attention outputs, the FFN catches intricate patterns that go beyond attention alone [1, 7].

3. CNNs vs. Vision Transformers: Comparative Analysis

3.1 Feature Extraction

CNNs use weight sharing and pooling to establish spatial hierarchies in order to learn hierarchical local filters. While deeper layers record object portions, earlier layers gather edges and textures [6, 7, 12, 15]. Strong inductive biases (locality, translation equivariance) make them data-efficient and robust for training [6, 15]. By segmenting images into patches and use self-attention to connect each patch to every other patch, ViTs offer direct global context access at every layer [1, 7, 9]. Compared to CNNs, which discard detail for class-discriminative features, ViTs retain more fine-grained detail; representations are consistent across depth and strongly propagate low-level features via residuals [1, 2, 7].

By adding convolutions to ViTs (CMT, Conformer, CEFormer) to add local inductive bias while retaining global attention, hybrid models consistently increase feature quality when compared to pure CNNs and pure ViTs [3, 11, 13, 20].

Aspect	CNN	Vision Transformer
Primitive Operations	Convolution + Pooling	Self-Attention + MLP
Inductive Bias	Strong (locality, translation)	Weak; must be learned
Global Context	Gradual via receptive field growth	Immediate via attention
Detail Retention (deep layers)	Lower, class-focused	Higher, rich low-level details

3.2 Local vs. Global Feature Learning

Local (CNN strength): Through tiny kernels, CNNs are particularly good at fine local structures, edges, tiny lesions, and textures [6, 12, 15]. particularly useful for applications like dentistry and small-object imaging that rely heavily on local cues and small datasets [6, 10, 17]. ViTs can naturally simulate long-range dependencies since global (ViT strength) self-attention aggregates throughout the entire image early on [1, 7, 9]. CNNs rely on local salient regions, while ViTs rely on broad geographical context, according to SHAP/LIME interpretability tools [9]. CNNs eliminate spatial detail for class separation, but ViTs retain it through depth [2]. Convolutional local streams and transformer global streams are fused in hybrid systems, such as ConViT, CMT, Conformer, and CVTrack, which perform exceptionally well in situations requiring wetland mapping, medical staging, and multi-region reasoning [3, 4, 14, 20].

3.3 Computational Efficiency

CNNs use fixed kernels and parameter sharing and are fast and lean, especially at small-to-medium resolution [6, 7, 15]. Larger datasets and careful optimization are required since Vanilla ViTs have $O(N^2)$ complexity due to full self-attention, which increases FLOPs and memory at high resolution [7, 9, 10]. CNNs perform better in terms of latency and efficiency in lightweight and small-data workloads, while ViTs perform better in high-data, high-compute regimes [7, 9, 10, 15]. CMT-S achieves 83.5% ImageNet top-1 with hybrid efficiency, which is 2× more efficient than EfficientNet and 14× more FLOP-efficient than DeiT [3]. While CNNs with CNN inductive bias perform well with fewer examples, ViTs need substantial pretraining to be computationally efficient per FLOP [1, 7, 9, 19].

3.4 Performance in Image Recognition

Vision Transformers (ViTs) have demonstrated robust performance in a variety of domains, frequently matching or outperforming CNNs when trained on massive datasets such as ImageNet, with improved scalability as data size grows [1, 7, 9, 10]. While ViTs usually achieve higher accuracy and resilience with sufficient data, CNNs are still more effective in low-data situations [7, 9, 10]. ViTs outperform models such as EfficientNet, ResNet, and VGG in face recognition in terms of accuracy, resistance to occlusions, and memory efficiency [5]. CNNs are still helpful as reliable baselines when data is limited, although ViT-based models generally perform better in medical and dental imaging when pretraining is sufficient [6, 17].

3.5 Hybrid CNN–Transformer Architectures

3.5.1 Motivation

Pure CNNs are good at local pattern extraction, but they struggle to handle long-range dependencies. Pure transformers can capture global relations even though they lack local priors and require large pretraining sets. Hybrid architectures combine localized injection with global attention to enhance generalization, sample efficiency, robustness to occlusion, and variable size [3, 6, 11, 20].

3.5.2 Integration Patterns

Hybrid vision models combine CNNs and Transformers to capture both local and global features. A CNN backbone (such ResNet or EfficientNet) is used to extract multi-scale feature maps, which are then converted into tokens for Transformer encoders. Some systems use an early convolutional stem with hierarchical Transformers for multi-scale learning, while others run CNN and Transformer branches in parallel and merge their features. Because smaller patches save more data but demand more processing power, tokenization is crucial. A higher learning rate for Transformer layers and a lower learning rate for the backbone are used to enhance the integrated model after the CNN is typically pretrained during training (for instance, on ImageNet).

3.5.3 Benefits in Medical

Hybrid CNN–Transformer architectures offer significant benefits in medical imaging by effectively combining local and global feature knowledge. CNNs enhance local lesion sensitivity by capturing fine-grained micro-textures and edge features, which are crucial for detecting minute abnormalities like lesions or microcalcifications. Transformers simultaneously represent connections between picture slices or patches to provide global context awareness. This makes it possible to assess tumor size and interactions with surrounding tissue more accurately, which is essential for multi-region analysis and accurate staging [6, 20]. Empirical research indicates that these hybrid models perform well in both radiology and histopathology tasks because they integrate precise patch-level information with more broad slide-level context. Clinically, this improves interpretability through attention mapping, increases false-positive control, and increases generalization across different imaging devices and scanners [6].

Attribute	CNN Only	Hybrid CNN–Transformer	ViT Only
Local detail	Excellent	Excellent	Good
Global context	Limited	Strong	Strong
Data needed	Low–moderate	Moderate	High
Compute	Low	Moderate	High
Medical imaging fit	Good	Best	Good (large pretraining)

IV. Literature Review

Despite its success in large-scale classification, the original ViT had a number of shortcomings, such as the requirement for very large training datasets (ImageNet-21K, JFT-300M), the high computational cost due to global self-attention, and the lack of hierarchical feature extraction [7, 12]. Several improved versions address these shortcomings by domain-specific adaptations, efficient training methods, and architectural modifications.

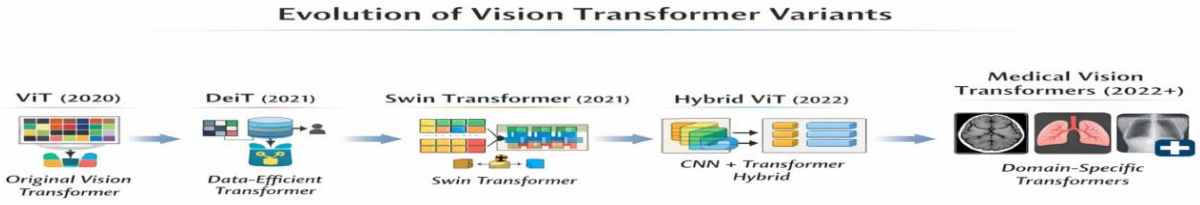

Figure 7: Evolution of Vision Transformer Variants

4.1 Evolution of ViT Models

Model	Key Contribution	Main Advantage
ViT	First Transformer-based vision model [7]	High accuracy on large datasets
DeiT	Knowledge distillation training	Works well with smaller datasets
Swin Transformer	Hierarchical windowed architecture	Efficient for large/medical images
Hybrid ViT	CNN + Transformer integration [3, 11]	Best of both architectures
Medical ViT	Domain-specific adaptation [6, 12]	Superior medical imaging performance

(1) Hierarchical feature representation: Swin Transformer presents multi-stage topologies that capture features at various scales, resembling CNN feature pyramids [12]; (2) Robustness, domain-specific transformers include specialized training techniques for noisy/heterogeneous medical datasets; (3) Data efficiency, knowledge distillation (DeiT), self-supervised learning, and hybrid designs reduce data requirements; and (4) Efficient attention, localized attention windows replace quadratic full self-attention [6, 12].

4.2 Swin Transformer

One of the most significant ViT variations is the Swin Transformer (Shifted Window Transformer) [12]. In contrast to the original ViT, it uses a hierarchical design akin to CNN feature pyramids to process pictures through several stages with gradually decreasing spatial resolution. Each level reduces complexity from $O(N^2)$ to $O(N)$ by using Transformer blocks that operate on local, non-overlapping windows instead of the full image. Attention Shifted to the Window: The defining mechanism. Cross-window information sharing without global attention is made possible by the shifting of windows between successive tiers. Benefits include enhanced feature representation, increased scalability, and decreased computing complexity. Because of their hierarchical efficiency, Swin Transformers are especially well-suited for high-resolution imaging and perform exceptionally well in image classification, object identification, semantic segmentation, and medical image analysis [12].

4.3 DeiT (Data-Efficient Image Transformer)

DeiT addresses the primary ViT problem of requiring huge datasets [7, 12]. It introduces knowledge distillation, where a CNN-based teaching paradigm directs transformer training. A special distillation token that learns from the teacher's soft labels rather than ground-truth labels is part of the patch sequence. DeiT can now outperform JFT-300M on ImageNet-1K alone thanks to this. The key advantages include lower data requirements, faster convergence, and higher training efficiency.

Feature	ViT	DeiT
Training Dataset	Very large (JFT-300M)	Medium-sized (ImageNet-1K)
Training Efficiency	Lower	Higher
Knowledge Distillation	No	Yes (distillation token)

4.4 Hybrid Vision Transformers

DeiT addresses the primary ViT problem of requiring huge datasets [7, 12]. It introduces knowledge distillation, where a CNN-based teaching paradigm directs transformer training. A special distillation token that learns from the teacher's soft labels rather than ground-truth labels is part of the patch sequence. DeiT can now outperform JFT-300M on ImageNet-1K alone. The key advantages include reduced data requirements, faster convergence, and increased training efficiency.

4.5 Medical-Specific Transformers

Complex anatomical features, high-resolution 3D pictures, and a paucity of labeled data present unique challenges in medical imaging [6]. Medical-specific transformer models combine self-supervised pretraining, multi-scale feature extraction, hybrid CNN-Transformer architectures, and transfer learning from natural image datasets to address these [6, 12].

Model	Application	Imaging Modality
TransUNet	Medical segmentation (U-Net + Transformer)	CT, MRI
UNETR	Volumetric organ segmentation	MRI
MedViT	Multi-modal disease detection	Multi-modal imaging

These models have demonstrated strong performance in tumor detection, brain MRI segmentation, and COVID-19 detection from chest CT scans [6, 12].

4.6 Adaptations for Domain-Specific Applications

ViTs have also been adapted for remote sensing (hierarchical transformers handle extremely high-resolution satellite images) and industrial inspection (attention-based localization finds small production faults). Common elements of domain-specific models include multi-scale feature extraction, unique preprocessing pipelines, and training methods based on domain data attributes [12, 18].

V. Conclusion

The Vision Transformer architecture represents a paradigm shift in computer vision by replacing convolutional processes with attention-based techniques. ViTs partition images into patches and treat them as sequences to efficiently model both local and global dependence. Strong feature representations that enable complex visual tasks include patch embedding, positional encoding, self-attention, multi-head attention, and feed-forward layers [1, 7, 12]. The comparison analysis presents a complicated picture. Because CNNs provide efficient, data-friendly local feature hierarchies, they remain the obvious choice in clinical settings with limited data and resources [6, 7, 15]. ViTs are superior at global context modeling and offer scalable performance with massive data and computing resources [1, 9]. The optimal architecture for medical imaging is continuously demonstrated to be hybrid CNN–ViT designs, which successfully combine convolutional inductive bias with transformer self-attention to balance feature quality, durability, and computational efficiency [3, 6, 20]. The Swin Transformer's hierarchical windowed attention removes the quadratic complexity bottleneck across ViT variants, which is advantageous for high-resolution CT and MRI images [12]. DeiT's knowledge distillation tackles data hunger [7]. Medical-specific models such as TransUNet and UNETR demonstrate the successful integration of general-purpose vision models with clinical image processing requirements through careful architectural adaption and domain-aware training strategies [6, 12]. As research advances, Vision Transformers are expected to play an increasingly significant role in clinical decision support, disease detection, and medical diagnosis.

REFERENCES

[1] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[2] R. Shi et al., "Visualization Comparison of Vision Transformers and Convolutional Neural Networks," IEEE Transactions on Multimedia, vol. 26, pp. 2327–2339, 2024.
[3] J. Guo et al., "CMT: Convolutional Neural Networks Meet Vision Transformers," Proc. IEEE/CVF CVPR, 2022.
[4] S. H. A. Moghaddam et al., "Integrating Local and Global Features in a Convolutional Vision Transformer for Wetland Mapping," IEEE Sensors Journal, vol. 25, pp. 8674–8683, 2025.
[5] M. Rodrigo et al., "Comprehensive Comparison Between Vision Transformers and CNNs for Face Recognition Tasks," Scientific Reports, vol. 14, 2024.
[6] S. Takahashi et al., "Comparison of Vision Transformers and CNNs in Medical Image Analysis: A Systematic Review," Journal of Medical Systems, 2024.
[7] J. Mauricio et al., "Comparing Vision Transformers and CNNs for Image Classification: A Literature Review," Applied Sciences, 2023.
[8] A. Qubadris et al., "An Overview Comparison Between Convolutional Networks and Vision Transformers," Proc. 7th Int. Conf. on Networking, Intelligent Systems and Security, 2024.
[9] M. K. Pasupuleti et al., "Vision Transformers vs. CNNs: Benchmarking Across Tasks," International Journal of Academic and Industrial Research Innovations (IJAIRI), 2025.

[10] B. Ali et al., "Vision Transformers in Image Restoration: A Survey," Sensors (Basel), 2023.
[11] T. Zhang et al., "Depth-Wise Convolutions in Vision Transformers for Efficient Training on Small Datasets," Neurocomputing, p. 128998, 2024.
[12] A. Khan et al., "A Survey of the Vision Transformers and Their CNN-Transformer Based Variants," Artificial Intelligence Review, 2023.
[13] L. Yin et al., "Convolution-Transformer for Image Feature Extraction," Computer Modeling in Engineering & Sciences, 2024.
[14] J. Wang et al., "CVTrack: Combined CNN and Vision Transformer Fusion Model for Visual Tracking," Sensors (Basel), 2024.
[15] O. Moutik et al., "CNNs or Vision Transformers: Who Will Win the Race for Action Recognition in Visual Data?" Sensors (Basel), 2023.
[16] M. C. Arslanoğlu et al., "Vision Transformers Versus CNNs: Comparing Robustness by Exploiting Varying Local Features," IEEE Access, vol. 13, pp. 65232–65245, 2025.
[17] T. Felek et al., "Evaluating Vision Transformers and CNNs in Dental Image Processing: A Systematic Review," BMC Oral Health, vol. 25, 2025.
[18] O. Elharrouss et al., "ViTs as Backbones: Leveraging Vision Transformers for Feature Extraction," Information Fusion, vol. 118, p. 102951, 2025.
[19] X. Li et al., "MSViT: Training Multiscale Vision Transformers for Image Retrieval," IEEE Transactions on Multimedia, vol. 26, pp. 2809–2823, 2024.
[20] Z. Peng et al., "Conformer: Local Features Coupling Global Representations for Recognition and Detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 45, pp. 9454–9468, 2023.